УДК 004

УВЕЛИЧЕНИЕ РОБАСТНОСТИ НЕЙРОННЫХ СЕТЕЙ ЗА СЧЕТ ГЕНЕРАЦИИ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ, ИНВАРИАНТНЫХ К АТРИБУТАМ

М. Р. Газизов[1], **К. А. Григорян**[2]

Казанский (Приволжский) федеральный университет, Казань

[1]gazizovmarat@gmail.com, [2]karigri@yandex.ru

Аннотация

Робастность модели к незначительным отклонениям в распределении исходных данных является важным критерием во многих задачах. Нейронные сети могут показывать высокую точность (accuracy) на обучающей выборке, но при этом качество на тестовой выборке может сильно падать из-за разного распределения данных, причем ситуация только усугубляется на уровне подгрупп внутри каждой категории.

В данной статье мы показываем, как робастность модели на уровне подгрупп может быть значительно улучшена с помощью подхода, основанного на доменной адаптации векторных представлений. Мы обнаружили, что применение состязательного подхода к ограничению векторных представлений дает существенный прирост метрики точности (accuracy) в сложной подгруппе по сравнению с предыдущими моделями. Метод протестирован на двух независимых наборах данных, точность в сложной подгруппе на наборе данных Waterbirds составляет 90.3 $\{y : waterbirds; a : landbackground\}$, а на наборе данных CelebA – 92.22 $\{y : blondhair; a : male\}$.

Ключевые слова*: робастная классификация, классификация изображений, генеративно-состязатель сети, доменная адаптация.*

ВВЕДЕНИЕ

В процессе обучения нейронных сетей в общем случае пытаются минимизировать среднее значение целевой функции ошибки на обучающей выборке или, другими словами, минимизируют эмпирический риск [1]. При этом мини-

мизация средней оценки целевой функции справедлива в случаях, когда тестовая выборка получена из распределения, независимого или идентичного обучающей выборке. В подобных задачах современные модели показывают высокое качество [2], но при этом проваливаются на примерах, редких и нетипичных для обучающей выборки [3–5]. Часто это делает невозможным применение моделей в реальной жизни, если эти модели используются для принятия решений [6] или оценки рисков [7].

К примеру, модель может полагаться на ложные корреляции атрибутов изображений *с целевой категорией* из-за неравномерного распределения значений атрибутов внутри категории, как в случае с метками на коже (атрибут) и наличием меланомы (целевая категория) у человек. На рисунке 1 показано неравномерное распределение меток между категориями злокачественных и доброкачественных образований, т. е. метки встречаются от 4.5 до 6 раз реже в категории со злокачественным образованием, чем в других категориях. Это приводит к увеличению доли ложно-отрицательных прогнозов в группе людей со злокачественными образованиями и метками на коже [8]. В этом и подобных случаях средняя оценка целевой функции не будет достоверно отображать устойчивость модели к наличию данных меток, а качество в сложной подгруппе может значительно отличаться в меньшую сторону от среднего по всей категории.

Рисунок. 1. Примеры изображений из набора данных ISIC [9–11] с метками и распределение меток среди категорий. Меланома – злокачественное образование. Себорейная кератома – доброкачественное образование. Последний столбец указывает на отсутствие образований.

Таким образом, требуется следить не только за средней оценкой показателя качества на уровне категории, но также замерять качество на уровне каждой из подгрупп внутри категории.

В данной статье мы изучаем подход, основанный на генерации инвариантных к атрибутам векторных представлений с помощью состязательных сетей на примере задачи классификации изображений. Мы предлагаем использовать доменную адаптацию для сближения векторных представлений примеров из разных подгрупп так, чтобы они становились неотличимы друг от друга на уровне категории, но все еще отличались на уровне целевого класса. Мы показываем, как адаптация векторных представлений может быть применима для минимизации зазора показателей качества между подгруппами, что значительно увеличивает устойчивость модели во время применения на тестовых наборах данных.

Предложенное нами решение показывает большую точность (accuracy) в сложных подгруппах на двух модельных наборах данных по сравнению с другими методами. Достигнутая нами точность в сложной подгруппе для набора данных Waterbirds составляет 90.3 {y : *waterbirds*;a : *landbackground*}, а для набора данных CelebA – 92.22 {y : *blondhair*;a : *male*}.

ОБЗОР ЛИТЕРАТУРЫ

Group distributionally robust optimization – GDRO [12]

Задача увеличения робастности модели может быть рассмотрена на уровне процесса оптимизации. В стандартном подходе весовые коэффициенты для разных подгрупп являются одинаковыми. Мы можем влиять на них явно через задание разных весов для разных групп или неявно через семплирование данных. Авторы этой статьи предлагают подход, основанный на динамической оптимизации весовых коэффициентов для каждой подгруппы во время процесса оптимизации модели. Также ими обнаружено, что увеличение регуляризации крайне важно для увеличения точности в сложной подгруппе. Они показывают теоретическую сходимость задачи оптимизации на выпуклых задачах. Помимо этого, ими предложен новый набор данных Waterbirds для сравнения качества работы разных моделей.

Class-conditional Learned Augmentations for Model Patching – CLAMP [13]

Задача доменной адаптации может решаться как на уровне векторных представлений, так и на уровне входных данных. Авторы данного подхода предлагают использовать генеративно-состязательные сети для трансформации исходных данных между подгруппами с сохранением целевого класса. Они исследуют подход, основанный на архитектуре нейронной сети CycleGAN [14]. Такая сеть позволяет решать задачу image-to-image translation без наличия спаренных изображений, в отличие от большинства методов, которые требуют наличия пар изображений из разных доменов [15, 16].

Для увеличения консистентности генерируемых изображений авторы предлагают использовать дополнительный модуль, задача которого – приближение синтетических данных к реальным данным.

Из положительных сторон генеративного подхода можно выделить его обобщающую способность, т. к. при наличии большого объема данных и достаточных по емкости моделей можно создавать данные, максимально близкие к данным из реального распределения [17, 18], тем самым устраняя дисбаланс между подгруппами.

Из минусов данного подхода можно выделить сложность обучения генеративных моделей [19] и высокие требования к вычислительным мощностям [18].

МЕТОДОЛОГИЯ

В данном разделе мы описываем предлагаемый состязательный подход к увеличению робастности модели (adversarial approach to increase model robustness – AAIMR).

Пусть дана выборка $x \in X$, целевая переменная y $\in Y$ и атрибут $a \in A$. Наша цель – обучить классификатор $f: C(F(x)) \to y$, который будет устойчив к изменению значения атрибута $a \in A$. При этом значение атрибута a известно только при обучении модели, но неизвестно во время ее применения. Таким образом, алгоритм не должен напрямую зависеть от a.

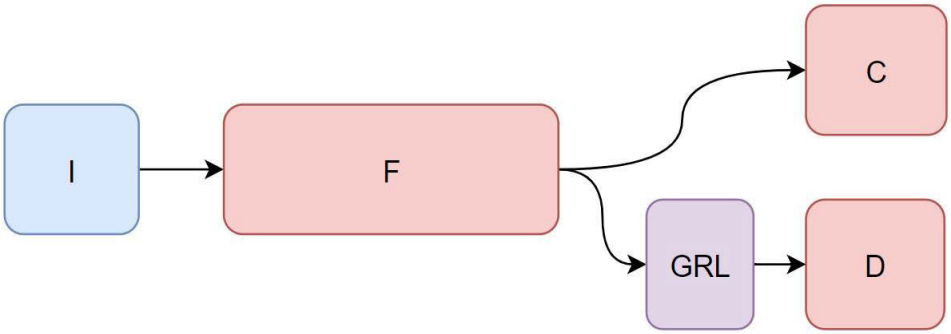


Рисунок 2. F – модель, преобразующая изображение в некоторое векторное представление, C – голова для предсказания значений целевого класса, D – дискриминатор для предсказания значений атрибута, GRL – слой обратных градиентов.

Очевидно, что переобучение модели C на значениях атрибутов a возможно только в том случае, если вектор $e = F(x)$ содержит информацию, позволяющую обучить такой классификатор $f: D(e) \to a$. Следовательно, если информация об атрибуте a не может быть получена из вектора e, то модель C будет инвариантна к значению атрибута a.

Таким образом, мы хотим поставить состязательную минимакс-задачу следующим образом: D обучается распознавать значение атрибута a, C – значение атрибута y. Задача F – научиться генерировать такие векторные представления e, чтобы дискриминатор D не мог правильно восстановить значение атрибута исходного изображения, при этом C мог правильно восстановить значение целевого класса. Для этого требуется оптимизировать следующий функционал:

$$E(\theta_f, \theta_c, \theta_d) = L\big(C\big(F\big(x\big|\theta_f\big)\big|\theta_c\big), y\big) - \lambda\, L\big(D\big(F\big(x\big|\theta_f\big)\big|\theta_d\big), a\big), \qquad (1)$$

$$\big(\hat{\theta}_f, \hat{\theta}_c\big) = E\big(\theta_f, \theta_c, \hat{\theta}_d\big) \underset{\theta_f, \theta_c}{\longrightarrow} \min, \quad \hat{\theta}_d = E\big(\hat{\theta}_f, \hat{\theta}_c, \theta_d\big) \underset{\theta_d}{\to} \max,$$

где L – целевая функция (в нашем случае – перекрестная энтропия), $\theta_f, \theta_c, \theta_d$ – оптимизируемые параметры моделей, λ – весовой коэффициент.

Чтобы функционал (1) можно было оптимизировать градиентным спуском с помощью метода обратного распространения ошибки, мы добавляем в архитектуру начальной сети слой обратных градиентов [20], как показано на рисунке 2.

В случае сходимости модель F будет продуцировать векторные представления, инвариантные к значению атрибута a, и, следовательно, модель C также будет инвариантна к изменению атрибута a.

Для обучения была взята модель ResNet-50 [21] с предобученными весами на наборе данных ImageNet [22]. В качестве оптимизатора использовался Adam [23]. Политика изменения скорости обучения не использовалась, скорость обучения была константной на всем протяжении процесса оптимизации.

Модель и другие параметры были выбраны в соответствие с [12] для возможности прямого сравнения результатов.

Для задачи Waterbirds [12] параметры оптимизации были зафиксированы weights decay = 1.0 и learning rate = 0.00001, для CelebA [20] weights decay = 0.1 и learning rate = 0.00001. Размер батча был зафиксирован на значении 128 для обеих моделей. Примеры были равномерно взвешены на уровне подгрупп. Весовой коэффициент слоя обратных градиентов λ=0.5.

Мы также дополнительно исследовали верхнюю оценку точности на задаче Waterbirds [12], для этого использовали маски, размеченные вручную. По маске вырезали объекты и зануляли фон, дальше обучали модель ResNet-50 [21] с равномерным взвешиванием примеров на уровне категорий, остальные параметры остались без изменений.

РЕЗУЛЬТАТЫ

Результаты, представленные ниже, получены на тестовых наборах данных для лучшей модели. В качестве таковой бралась та, которая имеет точность на всем валидационном наборе данных выше 0.9 и максимальную точность в сложной подгруппе.

Результаты, представленные в Таблице 1 для двух модельных задач, показывают, что предложенное решение дает значимое увеличение метрики качества в сложной группе при сохранении общей точности на том же уровне по сравнению с предыдущими подходами. Помимо этого, зазор между робастной точностью и общей точностью минимален.

Таблица 1. Результаты сравнения моделей. Общая точность считается на всей тестовой выборке. Робастная точность – это точность (accuracy) на сложной подгруппе

	CelebA		Waterbirds	
	Общая точность	Робастная точность	Общая точность	Робастная точность
ERM	94.8	41.1	97.3	60
GDRO [12]	91.8	88.3	93.2	86
CLAMP [13]	92.9	83.9	90.89	89.12
AAIMR	91.74	**92.22**	90.51	**90.03**

Таблица 2: Верхняя робастная оценка для набора данных Waterbirds. Общая точность считается на всей тестовой выборке. Робастная точность – это точность (accuracy) на сложной подгруппе

	Waterbirds	
	Средняя оценка	Робастная оценка
ERM	90.29	89.25

Результаты верхней оценки модели классификации на наборе данных Waterbirds показывают, что в данной задаче с большой вероятностью прирост точности за счет устранения корреляции между атрибутами и целевой переменной исчерпывается на уровне, указанном в Таблице 2, и дальнейшее улучшение связано с другими факторами.

ОБСУЖДЕНИЕ

Мы проанализировали подход доменной адаптации применительно к адаптации атрибутов изображений. Инвариантные к атрибутам признаки дают существенный прирост точности модели в сложной подгруппе и, как следствие, увеличивают робастность модели.

Наши результаты показывают, что доменная адаптация – перспективное направление для улучшения робастности модели.

В дальнейшем более глубокий анализ методов доменной адаптации, трансформации изображений между доменами и разделения векторных пространств может открыть немало интересных подходов к повышению робастности моделей.

СПИСОК ЛИТЕРАТУРЫ

1. *Vladimir Vapnik*. Principles of risk minimization for learning theory // Advances in Neural Information Processing Systems. 1992. P. 831–838.

2. *Christian Szegedy at all.* Inception-v4, inception-resnet and the impact of residual connections on learning. Thirty-first AAAI Conference on Artificial Intelligence, 2017.

3. *Dirk Hovy, Anders Søgaard*. Tagging performance correlates with author age // Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (volume 2: Short papers). 2015, P. 483–488.

4. *Nicole Shadowen*. Ethics and bias in machine learning: A technical study of what makes us “good”. The Transhumanism Handbook. Springer, 2019. P. 247–261.

5. *Osonde A Osoba, William Welser IV*. An intelligence in our image: The risks of bias and errors in artificial intelligence. Rand Corporation, 2017.

6. *Shai Danziger, Jonathan Levavu, Liora Avnaim-Pesso*. Extraneous factors in judicial decisions // Proceedings of the National Academy of Sciences 108.17 (2011). P. 6889–6892.

7. *Amitabha Mukerjee at all.* Multi-objective evolutionary algorithms for the risk-return trade-off in bank loan management // International Transactions in Operational Research 9.5. 2002. P. 583–597.

8. *Julia K. Winkler at all.* Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition // JAMA Dermatology 155.10. 2019. P. 1135–1141.

9. *Philipp Tschandl, Cliff Rosendahl, Harald Kittler*. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions // Scientific Data 5. 2018. P. 180161.

10. *Noel CF Codella* at all. Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC)// 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE. 2018. P. 168–172.

11. *Marc Combalia at all.* Bcn20000: Dermoscopic lesions in the wild // arXiv preprint arXiv:1908.02288, 2019.

12. *Shiori Sagawa at all.* Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization // arXiv preprint arXiv:1911.08731, 2019.

13. *Sharon Li Karan Goel Albert Gu, Chris R´e.* Automating the Art of Data Augmentation CLAMP: An Instantiation of Model Patching, 2020.
URL: http://hazyresearch.stanford.edu/data-aug-part-4.

14. *Jun-Yan Zhu at all.* Unpaired image-to-image translation using cycleconsistent adversarial networks // Proceedings of the IEEE international Conference on Computer Vision, 2017. P. 2223–2232.

15. *Phillip Isola at all.* Image-to-image translation with conditional adversarial networks // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. P. 1125–1134.

16. *Soumya Tripathy, Juho Kannala, Esa Rahtu*. Learning image-to-image translation using paired and unpaired training samples // Asian Conference on Computer Vision. Springer, 2018. P. 51–66.

17. *Ivan Anokhin at all.* High-Resolution Daytime Translation Without Domain Labels // arXiv preprint arXiv:2003.08791, 2020.

18. *Tero Karras at all.* Analyzing and improving the image quality of stylegan // arXiv preprint arXiv:1912.04958, 2019.

19. *Sangwoo Mo, Minsu Cho, Jinwoo Shin*. Instagan: Instance-aware imagetoimage translation // arXiv preprint arXiv:1812.10889, 2018.

20. *Yaroslav Ganin, Victor Lempitsky*. Unsupervised domain adaptation by backpropagation // arXiv preprint arXiv:1409.7495, 2014.

21. *Ying Tai, Jian Yang, Xiaoming Liu*. Image super-resolution via deep recursive residual network // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. P. 3147–3155.

22. *Jia Deng at all.* Imagenet: A large-scale hierarchical image database// 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE. 2009. P. 248–255.

23. *Diederik P Kingma, Jimmy Ba*. Adam: A method for stochastic optimization // arXiv preprint arXiv:1412.6980, 2014.

OF NEURAL NETWORK MODEL ROBUSTNESS THROUGH GENERATING INVARIANT TO ATTRIBUTES EMBEDDINGS

Marat Gazizov[1], Karen Grigoryan[2]

Kazan (Volga region) Federal University, Kazan

[1]gazizovmarat@gmail.com, [2]karigri@yandex.ru

Abstract

Model robustness to minor deviations in the distribution of input data is an important criterion in many tasks. Neural networks show high accuracy on training samples, but the quality on test samples can be dropped dramatically due to different data distributions, a situation that is exacerbated at the subgroup level within each category. In this article we show how the robustness of the model at the subgroup level can be significantly improved with the help of the domain adaptation approach to image embeddings. We have found that application of a competitive approach to embeddings limitation gives a significant increase of accuracy metrics in a complex subgroup in comparison with the previous models. The method was tested on two independent datasets, the accuracy in a complex subgroup on the Waterbirds dataset is 90.3 {*y* : *waterbirds*;*a* : *landbackground*}, on the CelebA dataset is 92.22 {*y* : *blondhair*;*a* : *male*}.

Keywords: *robust classification, image classification, generative adversarial networks, domain adaptation*

REFERENCES

1. *Vladimir Vapnik*. Principles of risk minimization for learning theory // Advances in Neural Information Processing Systems. 1992. P. 831–838.

2. *Christian Szegedy at all.* Inception-v4, inception-resnet and the impact of residual connections on learning. Thirty-first AAAI Conference on Artificial Intelligence, 2017.

3. *Dirk Hovy, Anders Søgaard*. Tagging performance correlates with author age // Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (volume 2: Short papers). 2015, P. 483–488.

4. *Nicole Shadowen*. Ethics and bias in machine learning: A technical study of what makes us "good". The Transhumanism Handbook. Springer, 2019. P. 247–261.

5. *Osonde A Osoba, William Welser IV*. An intelligence in our image: The risks of bias and errors in artificial intelligence. Rand Corporation, 2017.

6. *Shai Danziger, Jonathan Levav,u Liora Avnaim-Pesso*. Extraneous factors in judicial decisions // Proceedings of the National Academy of Sciences 108.17 (2011). P. 6889–6892.

7. *Amitabha Mukerjee at all.* Multi-objective evolutionary algorithms for the risk-return trade-off in bank loan management // International Transactions in Operational Research 9.5. 2002. P. 583–597.

8. *Julia K. Winkler at all.* Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition // JAMA Dermatology 155.10. 2019. P. 1135–1141.

9. *Philipp Tschandl, Cliff Rosendahl, Harald Kittler*. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions // Scientific Data 5. 2018. P. 180161.

10. *Noel CF Codella* at all. Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC)// 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE. 2018. P. 168–172.

11. *Marc Combalia at all.* Bcn20000: Dermoscopic lesions in the wild // arXiv preprint arXiv:1908.02288, 2019.

12. *Shiori Sagawa at all.* Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization // arXiv preprint arXiv:1911.08731, 2019.

13. *Sharon Li Karan Goel Albert Gu, Chris R´e.* Automating the Art of Data Augmentation CLAMP: An Instantiation of Model Patching, 2020. URL: http://hazyresearch.stanford.edu/data-aug-part-4.

14. *Jun-Yan Zhu at all.* Unpaired image-to-image translation using cycleconsistent adversarial networks // Proceedings of the IEEE international Conference on Computer Vision, 2017. P. 2223–2232.

15. *Phillip Isola at all.* Image-to-image translation with conditional adversarial networks // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. P. 1125–1134.

16. *Soumya Tripathy, Juho Kannala, Esa Rahtu*. Learning image-to-image translation using paired and unpaired training samples // Asian Conference on Computer Vision. Springer, 2018. P. 51–66.

17. *Ivan Anokhin at all.* High-Resolution Daytime Translation Without Domain Labels // arXiv preprint arXiv:2003.08791, 2020.

18. *Tero Karras at all.* Analyzing and improving the image quality of stylegan // arXiv preprint arXiv:1912.04958, 2019.

19. *Sangwoo Mo, Minsu Cho, Jinwoo Shin*. Instagan: Instance-aware imagetoimage translation // arXiv preprint arXiv:1812.10889, 2018.

20. *Yaroslav Ganin, Victor Lempitsky*. Unsupervised domain adaptation by backpropagation // arXiv preprint arXiv:1409.7495, 2014.

21. *Ying Tai, Jian Yang, Xiaoming Liu*. Image super-resolution via deep recursive residual network // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. P. 3147–3155.

22. *Jia Deng at all.* Imagenet: A large-scale hierarchical image database// 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009. P. 248–255.

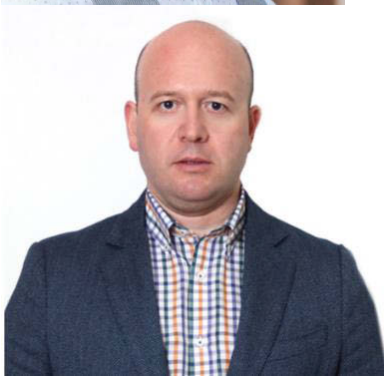
23. *Diederik P Kingma, Jimmy Ba*. Adam: A method for stochastic optimization // arXiv preprint arXiv:1412.6980, 2014.

СВЕДЕНИЯ ОБ АВТОРАХ

ГАЗИЗОВ Марат Рушанович – магистрант, Казанский (Приволжский) федеральный университет, г. Казань.

Marat Rushanovich GAZIZOV – graduate, Kazan (Volga region) Federal University, Kazan.

Email: gazizovmarat@gmail.com



ГРИГОРЯН Карен Альбертович – кандидат экономических наук, доцент, Казанский (Приволжский) федеральный университет, г. Казань.

Karen Albertovich GRIGORIAN – Candidate of Economics, Associate Professor, Kazan (Volga region) Federal University, Kazan.

Email: karigri@yandex.ru

Материал поступил в редакцию 4 июня 2020 года