Синтез изображений дорожных знаков с помощью условных порождающих противоборствующих нейросетей

П.В. Хрушков[1], В.И. Шахуро[1,2], А.С. Конушин[1,2]
pvkhrushkov@edu.hse.ru|vlad.shakhuro@hse.ru|anton.konushin@graphics.msu.ru
[1]НИУ Высшая Школа Экономики, Москва, Россия;
[2]МГУ имени М. В. Ломоносова, Москва, Россия

В работе рассматривается метод генерации синтетических обучающих выборок для задачи классификации дорожных знаков. Метод основан на использовании порождающих конкурирующих нейросетей и метрики Васерштейна. Исследуется метод условной генерации изображений, когда на вход порождающей нейросети подается случайный шум и метка класса изображения, которое нужно сгенерировать. Для обучения такой нейросети предлагается использовать перекрестную энтропию в добавление к метрике Васерштейна. Для стабилизации процесса обучения используются веса для обучающей выборки. Экспериментальная оценка метода показывает, что условная порождающая сеть работает лучше, чем простая генерация дорожных знаков по иконке, однако не дотягивает до метода, в котором для каждого класса обучается отдельная порождающая нейросеть.

Ключевые слова: классификация дорожных знаков, синтетические выборки, условные порождающие противоборствующие нейросети

Generation of synthetic traffic sign images using conditional generative adversarial networks

P.V. Khrushkov[1], V.I. Shakhuro[1,2], A.S. Konushin[1,2]
pvkhrushkov@edu.hse.ru|vlad.shakhuro@hse.ru|anton.konushin@graphics.msu.ru
[1]NRU Higher School of Economics, Moscow, Russia;
[2]Lomonosov Moscow State University, Moscow, Russia

In this work we research method for synthesis of training samples for classification of traffic signs. Method is based on generative adversarial networks and Wasserstein metric. We consider conditional generative neural network, which takes random noise and class label as input and outputs image of object of needed class. To train such generative network, we use cross entropy in addition to Wasserstein metric. To stabilize training, we use weighting of samples. Experimental evaluation of method shows that conditional neural network outputs samples which are better than naive samples generated using icons, but worse than set of networks trained separately for each traffic sign class.

Keywords: traffic sign classification, synthetic data, conditional generative adversarial networks.

1. Введение

В настоящее время прогресс в компьютерном зрении в значительной степени обусловлен появлением больших наборов размеченных данных. В качестве примера таких наборов данных можно привести [3, 5, 11, 12] для задач классификации, сегментации и детектирования объектов на изображениях. В зависимости от сложности разметки такие наборы данных насчитывают от 5000 (Cityscapes, семантическая сегментация) до 9 миллионов (OpenImages, детектирование объектов) и 14 миллионов (ImageNet, классификация) изображений. Разметка таких выборок — дорогой и трудозатратный процесс, который можно провести силами большого количества наемных либо добровольных рабочих. При этом даже после составления размеченной выборки изображений могут возникнуть проблемы:

1. Существуют редкие классы объектов. Даже несмотря на большие размеры выборок, обучающих примеров изображений таких объектов недостаточно для обучения современных методов машинного обучения (как правило, нейросетевых).
2. При появлении новых классов объектов требуется обновление выборки.

Описываемую проблему получения качественных обучающих выборок можно решить с помощью генерации синтетических изображений. В последнее время стали активно исследоваться модели порождающих конкурирующих нейросетей для генерации реалистичных изображений. В этих моделях используются две нейросети: генератор и критик (используется подход и терминология из [1]). Генератор принимает на вход случайный шум (например, нормальный) и выдает изображение, а критик выдает меру реалистичности поданного ему на вход изображения. В данной работе исследуется генератор, который помимо случайного шума на вход принимают также метку класса изображения, которое нужно сгенерировать. Чтобы стабильно обучать такой генератор, в функцию потерь критика добавляется дополнительное слагаемое. Помимо меры реалистичности изображения он должен выдавать вектор вероятностей меток класса. Кроме этого, чтобы бороться с несбалансированностью обучающей выборки, задаются веса для обучающих приме-

The publication of the conference proceedings was supported by the RFBR, grant No. 18-07-20045\18

ров. Два перечисленных метода стабилизации обучения условной порождающей нейросети являются основным вкладом данной работы. Качество работы метода генерации синтетических изображений оценивается на задаче классификации дорожных знаков.

2. Обзор методов генерации синтетических изображений

Синтетические обучающие выборки применяются для обучения алгоритмов распознавания изображений в случаях, когда получить и разметить реальные данные невозможно или требуется слишком много ресурсов. Один из примеров такой ситуации — анализ медицинских изображений. Данные могут быть труднодоступны (снимки пациентов с редким заболеванием) и трудоемки в разметке (для сегментации снимков, например, может понадобиться большое количество времени высококвалифицированного специалиста). Генерация реалистичных изображений — плохо определенная задача, т.к. в обработке изображений и компьютерном зрении на сегодняшний день не существует метрики фотореалистичности изображения, а в каждой задаче обучающая выборка имеет особенности. Кроме этого, пока не придумано надежного способа сравнения между собой различных генераторов синтетических изображений [18]. Поэтому для генерации обучающих примеров сейчас используются два метода: размножение реальных данных и трехмерное моделирование.

Размножение изображений активно используется для обучения нейронных сетей. В [13] для обучения классификатора изображений обучающая выборка увеличивается на порядки за счет случайной обрезки изображения и зеркальных отражений относительно вертикальной оси изображения. В [2] изображения дорожных знаков размножаются за счет случайных поворотов, сдвигов и масштабирований.

Трехмерное моделирование для создания обучающих выборок успешно используется в задачах, где не требуется высокая степень фотореалистичности. Например, в [16] обучается регрессор позы человека по карте глубины. С помощью трехмерного моделирования генерируются зашумленные карты глубины, которые, в отличие от реалистичных RGB-изображений, сгенерировать достаточно просто. Для задачи вычисления оптического потока с помощью сверточных нейронных сетей [6] используется нереалистичная выборка «Летающие стулья». Для вычисления оптического потока нейросеть должна научиться сопоставлять области двух кадров, поэтому правдоподобность данных не требуется.

В [8, 15] игровые движки используются для генерации размеченных городских сцен. Экспериментальная оценка показывает, что только синтетических данных недостаточно для качественного обучения алгоритма детектирования объектов и сегментирования изображений. Однако использование синтетических данных вместе с реальными позволяет улучшить качество итогового алгоритма.

В работах [4, 14] рассматривается простой метод генерации изображений дорожных знаков. Знак — стандартизированный объект, поэтому для него в качестве модели можно взять иконку. Затем с помощью случайных преобразований (гауссово размытие, размытие движения, поворот, сдвиг, масштабирование, наложение на фон, изменение контраста и цветности) из иконки получается изображение дорожного знака. Метод требует априорного задания преобразований, применяемых к модели.

В последнее время появился и активно развивается метод генерации фотореалистичных изображений на основе порождающих конкурирующих нейронных сетей [9]. В этом подходе попеременно обучаются две нейросети: генератор и дискриминатор (критик). Генератор преобразует входной случайный шум в реалистичное изображение, а дискриминатор пытается отличить реальное изображение от сгенерированного. Генератор и дискриминатор обучаются попеременно. В качстве функции потерь используется бинарная перекрестная энтропия.

Конкурирующие порождающие нейронные сети также используются для генерации обучающих выборок. В [19] синтетические данные в добавление к реальным данным используются для повышения качества повторной идентификации людей в видео. В [7] синтетические данные добавляются в обучающую выборку для улучшения качества классификации поражений печени. В [20] порождающие нейронные сети используются для генерации изображений дорожных знаков, при этом для каждого класса обучалась собственная нейросеть-генератор. В настоящей работе ставится следующая цель: обучить одну условную нейросеть-генератор дорожных знаков, т.е. нейросеть, принимающую на вход метку класса, и синтезирующую изображение этого класса.

3. Описание метода условной генерации дорожных знаков

В данной работе используется метод обучения порождающих нейросетей с помощью метрики Васерштейна. В процессе обучения участвуют две нейросети: генератор g_θ и критик f_w. Сеть-генератор получает на вход нормальный шум и метку класса и генерирует изображение. Сеть-критик получает на вход изображение и пытается отличить реальное изображение от сгенерированного. Для обучения генератора и критика используется метрика Васерштейна:

$$W(\mathsf{P}_r, \mathsf{P}_g) = \max_{w \in \mathcal{W}} \mathsf{E}_{x \sim \mathsf{P}_r}[f_w(x)] - \mathsf{E}_{z \sim p(z)}[f_w(g_\theta(z))].$$

Здесь $z \sim p(z)$ — случайный шум, подаваемый на вход нейросети-генератору, P_r и P_g — распределения реальных и синтетических изображений соответственно.

Для того, чтобы метрику можно было использовать для обучения нейросетей, нужно добиться того, чтобы функция f_w была липшицевой с константой 1 (подроб-

Работа опубликована при финансовой поддержке РФФИ, грант 18-07-20045\18

нее об этом условии см. [1]). В работе [10] это достигается с помощью использования дополнительного слагаемого в функции потерь:

$$L_R = -\lambda(||\nabla_{\hat{x}} f_w(\hat{x})|| - 1)^2.$$

Здесь $\hat{x} = tx + (1-t)g_\theta(z)$, $z \sim p(z)$, $t \sim U[0;1]$ — выпуклая комбинация реального и синтезированного изображения.

Для того, чтобы генератор учитывал метку класса, критик дополнительно учится классифицировать сгенерированные изображения на заданное количество классов. Для этого используется перекрестная энтропия:

$$L_C = \mathsf{E}[\log P(C = c|I_{real})] + \mathsf{E}[\log P(C = c|I_{fake})]$$

Здесь c — метка класса реального или синтезированного изображения.

Итоговая функция потерь L — сумма $W(\mathsf{P}_r, \mathsf{P}_g)$, L_C и L_R с некоторыми весами.

4. Обучение на несбалансированной выборке

Напомним, что в функцию потерь входит слагаемое L_R, в котором используется выпуклая комбинация $\hat{x}$ синтетического и реального изображений. Заметим, что не имеет смысла находит выпуклую комбинацию изображений двух различных классов. Процесс обучения устроен следующим образом:

1. Сэмплируется мини-батч реальных изображений.
2. Генерируем синтетические изображения с теми же метками, что и изображения в мини-батче. Таким образом, для каждого реального изображения имеем синтетическое с тем же классом.

Отметим также, что на шаге обновления генератора нужно сэмплировать метки классов для генерации изображений. Важно здесь то, что необходимо сэмплировать метку класса с той вероятностью, с которой она встречается в обучающей выборке. Это видно из формулы полной вероятности:

$$\mathsf{P}_r = p(y_1)\mathsf{P}_r(x|y_1) + \cdots + p(y_k)\mathsf{P}_r(x|y_k).$$

Даже если генератор умеет хорошо синтезировать объекты по метке (т.е. хорошо оценил условные распределения вида $\mathsf{P}_r(x|y)$), но метки сэмплируются не с вероятностями $p(y_1), \ldots, p(y_k)$, то получается совершенно другое распределение, отличное от P_r. Это, в свою очередь, означает, что функция потерь штрафует генератор за «неправильно» выученное распределение.

Предположим теперь, что какая-то метка y_i встречается намного чаще другой метки y_j. Это означает, что вероятностная мера, соответствующая y_i, войдёт в P_r с большим весом, нежели вероятностная мера, соответствующая y_i. Из этого следует, что генератору выгоднее генерировать объекты с меткой y_i лучше, чем объекты с меткой y_j. Теоретически WGAN способен идеально оценить реальное распределение, однако наша генерирующая модель ограничена в своих возможностях и не может сохранить в себе слишком много информации. Поэтому если WGAN и будет стоять перед выбором: качеством генерации какого класса пожертвовать, этим классом будет скорее y_j, а не y_i.

Чтобы достичь равного качества генерации изображений разных классов, добьемся равномерности распределения классов. Естественно, можно было бы применить undersampling или oversampling и непосредственно сравнять число объектов каждого класса, однако можно поступить по-другому. Как уже было сказано, будем учить нейросеть оценивать не исходное распределение P_r, а его «равномерную версию»

$$\mathsf{P}'_r = \frac{\mathsf{P}_r(x|y_1) + \cdots + \mathsf{P}_r(x|y_k)}{k}.$$

Заметим, что

$$\mathsf{E}_{x\sim\mathsf{P}'_r}[f(x)] = \int_{\mathcal{X}} f(x)\mathsf{P}'_r(dx) = \int_{\mathcal{X}} \frac{f(x)}{k} \sum_{i=1}^{k} \mathsf{P}_r(dx|y_i)$$

$$= \int_{\mathcal{X}} \frac{f(x)}{k} \sum_{i=1}^{k} \frac{p(y_i)}{p(y_i)} \mathsf{P}_r(dx|y_i)$$

$$= \int_{\mathcal{X}} \sum_{i=1}^{k} \frac{f(x)}{kp(y_i)} p(y_i)\mathsf{P}_r(dx|y_i)$$

$$= \mathsf{E}_{x,y\sim\mathsf{P}_r}\left[\frac{f(x)}{kp(y)}\right],$$

что соответствует введению веса $\frac{1}{kp(y)}$ для объектов класса y. Таким образом, можно пересчитывать все математические ожидания в задаче оптимизации WGAN и заставить модель оценивать «равномерную версию» реального распределения.

5. Экспериментальная оценка

Для экспериментальной оценки предложенного метода использовалась база автодорожных знаков gtsrb [17]. Эта выборка содержит 52 тысячи изображений 43 классов дорожных знаков.

На рис. 1 показаны примеры изображений, генерируемых с помощью условной порождающей нейросети, обученной с помощью метрики Васерштейна и вспомогательного классификатора. Многие изображения получаются визуально неотличимыми от реальных, однако некоторые классы путаются друг с другом (например, ограничение скорости 20 и класс «осторожно, дети», ограничения скорости 100 и 120).

Для получения количественных оценок был обучен нейросетевой классификатор дорожных знаков. Архитектура нейросети взята из [2] и описана в табл. 1.

В табл. 2 показаны результаты классификатора (точность многоклассовой классификации в процентах), обученного на различных тренировочных выборках. Метод условной генерации дорожных знаков, предложенный в данной работе (строчки «условная WGAN» в табл. 2) позволяет генерировать более качественные изображения, чем метод генерации по иконке дорожного знака, однако проигрывает по качеству 43 нейросетям, отдельно обученным на разных классах (строчки «WGAN синтетика» в табл. 2).

The publication of the conference proceedings was supported by the RFBR, grant No. 18-07-20045\18

#	Тип	Кол-во карт и нейронов в слое	Ядро
0	Input	3 карты по 48×48 нейронов	
1	Conv	100 карт по 100×100 нейронов	7×7
2	Maxpool	100 карт по 21×21 нейронов	2×2
3	Conv	150 карт по 18×18 нейронов	4×4
4	Maxpool	150 карт по 9×9 нейронов	2×2
5	Conv	250 карт по 6×6 нейронов	4×4
6	Maxpool	250 карт по 3×3 нейронов	2×2
7	FC	300 нейронов	1×1
8	FC	43 нейрона	1×1

Таблица 1. Архитектура свёрточной нейросети, использовавшейся для классификации знаков.

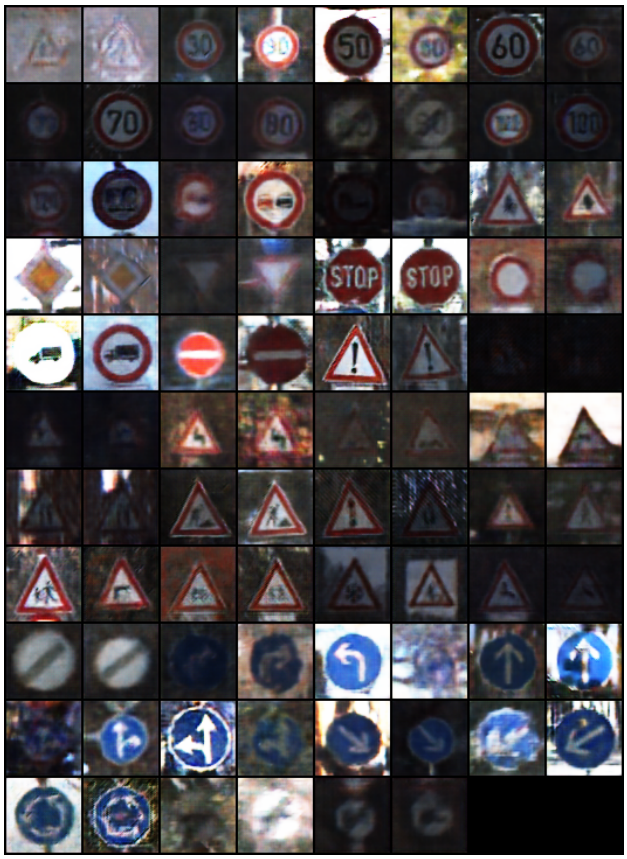

Рис. 1. Примеры изображений, сгенерированных с помощью условного Wasserstein GAN.

6. Заключение

В данной работе предложен метод обучения условной нейросети с помощью метрики Васерштейна. Обученная нейросеть получает на вход случайный шум и метку класса, а на выход выдает изображение заданного класса. Экспериментальная оценка метода на задаче классификации показала, что условная нейросеть работает лучше наивной генерации синтетических знаков по иконке, однако проигрывает по качеству поклассово обученным порождающим нейросетям.

7. Благодарность

Работа выполнена при поддержке гранта РНФ 17-71-20072 «Нейробайесовские методы в задачах машинного обучения, масштабируемой оптимизации и компьютерного зрения».

8. Литература

[1] Arjovsky M., Chintala S., Bottou L. Wasserstein gan //arXiv preprint arXiv:1701.07875. – 2017.

[2] Cireşan D. et al. Multi-column deep neural network for traffic sign classification //Neural networks. – 2012. – T. 32. – C. 333-338.

[3] Cordts M. et al. The cityscapes dataset for semantic urban scene understanding //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2016. – C. 3213-3223.

[4] Chigorin A., Konushin A. A system for large-scale automatic traffic sign recognition and mapping //CMRT13–City Models, Roads and Traffic. – 2013. – T. 2013. – C. 13-17.

[5] Deng J. et al. Imagenet: A large-scale hierarchical image database //Computer Vision and Pattern Recognition, 2009. - C. 248-255.

[6] Dosovitskiy A. et al. Flownet: Learning optical flow with convolutional networks //Proceedings of the IEEE International Conference on Computer Vision. – 2015. – C. 2758-2766.

[7] Frid-Adar M. et al. Synthetic data augmentation using GAN for improved liver lesion classification //Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on. – IEEE, 2018. – C. 289-293.

[8] Gaidon A. et al. Virtual worlds as proxy for multi-object tracking analysis //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2016. – C. 4340-4349.

[9] Goodfellow I. et al. Generative adversarial nets //Advances in neural information processing systems. – 2014. – C. 2672-2680.

[10] Gulrajani I. et al. Improved training of wasserstein gans //Advances in Neural Information Processing Systems. – 2017. – C. 5767-5777.

[11] Krasin I. et al. OpenImages: A public dataset for large-scale multi-label and multi-class image classification, 2017.

[12] Lin T. Y. et al. Microsoft coco: Common objects in context //European conference on computer vision. – Springer, Cham, 2014. – C. 740-755.

[13] Krizhevsky A., Sutskever I., Hinton G. E. Imagenet classification with deep convolutional neural networks //Advances in neural information processing systems. – 2012. – C. 1097-1105.

[14] Moiseev B. et al. Evaluation of traffic sign recognition methods trained on synthetically generated data //International Conference on Advanced Concepts for Intelligent Vision Systems. – Springer, Cham, 2013. – C. 576-583.

Работа опубликована при финансовой поддержке РФФИ, грант 18-07-20045\18

Тренировочная выборка	39 тыс. без размножения	215 тыс. без размножения	39 тыс. / 215 тыс. с размножением
Реальные данные	96.6	—	98.4 / —
WGAN синтетика	95.3	96.1	97.6 / 98.1
Реальные данные + WGAN синтетика	—	97.7	— / 98.4
Условная WGAN	79.2	83.7	81.3 / 81.5
Реальные данные + условная WGAN	—	95.2	— / 95.5
Синтетика по иконке	46.5	53.7	67.8 / 69.7
Реальные данные + синтетика по иконке	—	96.5	— / 97.9

Таблица 2. Результаты тестирования классификатора на различных выборках знаках

[15] Richter S. R. et al. Playing for data: Ground truth from computer games //European Conference on Computer Vision. – Springer, Cham, 2016. – С. 102-118.

[16] Shotton J. et al. Real-time human pose recognition in parts from single depth images //Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. – Ieee, 2011. – С. 1297-1304.

[17] Stallkamp J. et al. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition //Neural networks. – 2012. – Т. 32. – С. 323-332.

[18] Theis L., Oord A., Bethge M. A note on the evaluation of generative models //arXiv preprint arXiv:1511.01844. – 2015.

[19] Zheng Z., Zheng L., Yang Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro //arXiv preprint arXiv:1701.07717. – 2017. – Т. 3.

[20] Шахуро В. И., Конушин А. С. Синтез обучающих выборок для классификации дорожных знаков с помощью нейросетей //Компьютерная оптика. – 2018. – Т. 42. – №. 1.

The publication of the conference proceedings was supported by the RFBR, grant No. 18-07-20045\18