GAN-SSL Classification for Identification Expertise in Chemistry

Aleksandra Maksimova
Theory Control System Department
Institute of Applied Mathematics and Mechanics
Donetsk, DPR
maximova.alexandra@mail.ru

Abstract. **In this work we investigate the generative adversarial nets for classification problem of identification expertise in Chemistry. The identification expertise problem is challenging for classification because of complex structure of classes, outliers and cross-classes. The generative-adversarial nest for semi-supervised learning (GAN-SSL) is proposed for complex classification problem. The training samples are partially labeled for the semi-supervised tasks. Two groups of experiments were carried out. The first group of experiments for the model dataset that consist of classes of points normally distributed about vertices an eight-dimensional hypercube. The second groups of experiments for the petrol identification expertise dataset we get from laboratory of petrol quality. The experiments with good model examples get good quality more than 99%. The classification model for petrol identification expertise was created and has 93% quality but convergences training much worse. In this work we use GAN-SSL classification on petrol identification expertise example, but this classification model can be used for diesel fuel, household chemicals items, different oils and for various other objects.**

Keywords: **GAN, classification, identification expertise, semi-supervised learning**

I. INTRODUCTION

Proposed in 2014 by Jan Goodfellow [1] generative adversarial nets in [2] were improved for classification problems using semi-supervised learning. In semi-supervised learning is used unlabeled data for end-to-end learning of classifiers. The training samples are partially labeled for the semi-supervised tasks.

The identification expertise problem in Chemistry is formally seems as a classification problem [3]. In our previous works was proposed fuzzy portrait method for identification expertise of petrol [4], [5]. The quality of received models heavily dependent on available data. The specific problem for identification expertise is a very few of items for some classes, for example 15 items. If we consider objects of each class as elements of a certain probability distribution, we don`t have enough items to get good classifier. The representation of samples set is not good enough. The main idea is to use semi-supervised learning to overcome our limitations with datasets using generative adversarial nets for semi-supervised learning (GAN-SSL). The work is aimed at research of novel state-of-the-art GAN-SSL classification method.

II. RELATED WORKS

GAN publications have increasingly focused on the use of class labels. The first multiclass inference strategy for GAN was developed in [6], where the number of outputs of the discriminant classifier is equal to the number of classes, and training is carried out both on unmarked and partially marked data. Such a network is called categorical generative adversarial network (CatGan).

The most interesting for identification expertise classification problem is proposed in [2], [7] classification model. The number of outputs of the discriminator corresponds to the number of real classes and one more for the fake class, produces by generator. This strategy is good working for semi-supervised learning using the GAN loss functions.

In [8] proposed novel approach to semi-supervised learning on graphs – GraphSgan. Generator and classifier play a novel competitive game, when generator generates fake samples. This idea can help in identification expertise to find counterfeit items.

There are some papers where classical GANs architectures, like DCGAN and PGGAN using for classification [9]. The generator is training to produce realistic chest X-ray images and lymph node histology images. These images add to training data sets for classical convolutional net.

III. MODEL FRAMEWORK

A. Problem Definition

We consider petrol identification expertise as classification problem. We should identify the petrol producer and mark by a sample of petrol with solving the classification problem.

We have in our consideration ten classes of objects grouped by producer and mark parameters. We use eight features: research octane, motor octane number, density, volume fraction of olefinic hydrocarbons, benzene, toluene and mass fraction of methanol and of oxygen.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

The identification expertise problem belongs to sensitive to negative examples problem. The reason is the frequent cases of falsification of identification objects, when the composition of the product partially changes when using cheaper substitutes and additives.

B. GAN-SSL Architecture

We used GAN-SSL architecture proposed in [2]. GAN architecture includes two networks: generator and discriminator neural nets. The goal of model training is to train a generator $G(z)$ that produces samples from the data distribution $p_{data}(x)$ by transforming vector of noise z as $x = G(z)$. The discriminator is training to distinguish real data from the generator distribution $p_{data}(x)$. For GAN-SSL architecture discriminator is changed to standard K-classes classifier. We do semi-supervised learning with any standard classifier by adding samples from the GAN generator to data set and add $K+1$ class to classifier for these samples labelled like "generated fake".

Both discriminator and generator in GAN-SSL network are multiple layer perceptrons. The generator takes noise z from uniform distribution on the interval [0,1) as input and outputs fake samples having the similar shape as x. Batch normalization is used in generator [10]. It is used weight normalization trick [11] before output layer in generator. There are three layers in the generator.

Discriminator consists of six linear weight norm layers based on weight normalization trick too. It is used additive Gaussian noise in every layer before output for smoothing purpose in the training mode only.

The discriminator takes in object feature vector x as input and outputs K-dimensional vector of logits $\{l_1, \ldots, l_K\}$ and one more input for "generated fake" class. Then we can use softmax activation function to get class probabilities $p_{model}(y=j|x)=\exp(l_j)/\text{sum}(\exp(l_k))$. In practice, we only consider the first K outputs and assume the output for "generated fake" class is always 0 before softmax, because subtracting identical number from all units before softmax does not change the softmax results.

The number of neurons in hidden layers depends on identification expertise dataset and can be modified corresponding to velocity of problem. The discriminator must be more powerful than the generator and powerful enough for specific task. Denote the base number of neurons as N_{base} for experimental result section. It means, that there are N_{base} neurons in layers of generator and $2N_{base}$, $2N_{base}$, N_{base}, N_{base}, N_{base}, respectively in layers of discriminator.

The optimal discriminator in GAN-SSL is expected to be perfect on labeled and unlabeled data, but the generator will be always imperfect [12].

C. Learning Algorithm

There are two techniques to improve the training of GANs proposed in [2]. We use feature matching technique that addresses the instability of GANs by specifying a new objective for generator that avert it from overtraining on the current discriminator. We don`t use minibatch discrimination in our work because it further improves the generator examples that is not necessary for identification expertise problem.

The loss function for discriminator consists of two components: supervised $L_{supervised}$ and unsupervised $L_{unsupervised}$ loss functions [2]:

$$L_{supervised}=\mathbf{E}_{x,y\sim pdata(x,y)}\log[D(x)], \quad (1)$$

$$L_{unsupervised}=-\mathbf{E}_{x\sim pdata(x,y)}\log[D(x)] - -\mathbf{E}_{z\sim noise}\log[1-D(G(z))], \quad (2)$$

where $\mathbf{E}_{x,y\sim pdata(x,y)}$ is expectation of labeled data, $\mathbf{E}_{x\sim pdata(x,y)}$ is expectation of unlabeled data, $\mathbf{E}_{z\sim noise}$ is expectation of noise, $D(x)$ is output of discriminator applying softmax, $G(z)$ is output of generator.

The output of discriminator layer before softmax we denote as $\boldsymbol{f}(x)$ function that uses in new objective function for generator:

$$L_{gen}=||\mathbf{E}_{x\sim pdata(x,y)}\boldsymbol{f}(x) - \mathbf{E}_{z\sim noise}\boldsymbol{f}(G(z))||_{L2}, \quad (3)$$

where $\mathbf{E}_{x\sim pdata(x,y)}$ and $\mathbf{E}_{z\sim noise}$ like in (2), $||f||_{L2}$ is L_2 norm.

We use minibatch stochastic **Algorithm 1** to train generator and discriminator iteratively minimizing their losses.

Algorithm 1: Minibatch stochastic gradient descent training of GAN-SSL for identification expertise.

Input: $X^1_{unlabeled}$- identification objects dataset – shuffled unlabeled data set #1
$X^2_{unlabeled}$- identification objects dataset – shuffled unlabeled data set #2
$X_{labeled}$- identification objects dataset for supervised learning; y – labels of classes correspond to pare (producer, mark of petrol)

Make $X^1_{unlabeled}$, $X^2_{unlabeled}$, $X_{labeled}$ equal length datasets by folding $X_{labeled}$

for *number of epochs* ***do***
- Sample minibatch from $X^1_{unlabeled}$
- Sample minibatch from $X_{labeled}$
- Sample minibatch from noise prior $p_g(z)$
- Update the discriminator by descending gradients of losses:
  $$\boldsymbol{L=L_{supervised}+L_{unsuprvised}}$$
- Sample minibatch from $X^2_{unlabeled}$
- Sample minibatch from noise prior $p_g(z)$ *as* $G(z)$
- ***for*** *2 steps* ***do***

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

Update the discriminator by descending gradients of losses:
$\boldsymbol{L = L_{gen}}$

end for

end for

The gradient-based updates can use any standard gradient-based learning rule. We use Adam, based on adaptive estimates of lower-order moments [13].

We train the model for different power of the generator and discriminator adjusting base neuron parameter N_{base}.

IV. Experimantal Results

We use GAN-SSL for two classification problems with the same numbers of features and classes to compare and analyze the speed and quality of training GAN-SSL. First one is model examples, but second is real petrol identification expertise problem with cross-classes and outliers.

The model data set consist of classes of points normally distributed about vertices of eight-dimensional hypercube. These data have the same objects in every class and easy for classification. The total number of objects is 5000, of which 100 labeled examples of an equal number from each class.

The visualization of training for experiments with model dataset is presented in Fig. 1. The x-axis shows the number of epochs. The model 1 dataset consists of 10 normally distributes classes, and the model 2 consist of classes with two normally distributed clusters for every class. The 1% of object is outliers for every models. The number of neurons N_{base} in the third example "weak net" is 15, in the rest – 25.

The petrol identification expertise data we get from laboratory of petrol quality. There are different number of objects in classes from 72 in smallest to 671 in the most popular. We use only 100 labeled examples for training. The total number of objects is 6710, of which 100 labeled examples of an equal number from each class.

We train three GAN-SSL networks with different N_{base} parameter for network architecture: 25, 35 and 45. The visualization of training is presented in Fig. 2. The x-axis shows the number of epoch.

The experiments with good model examples get good quality more than 99%. The training process for the model data was fairly stable, without sharp fluctuations. Expectedly, training was fastest on model 1 data. The GANs with N_{base} equal to 25 train better than for weak net, where the base number of neurons is equal to 15. The generator loss L_{gen} grows corresponding to imperfect quality of generator, which was substantiated in [12].

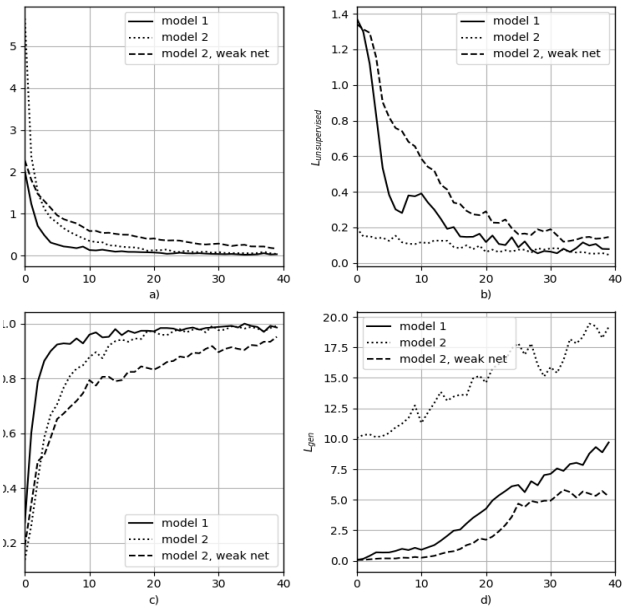

Fig. 1. Visualisation of training process for model datasets: simple model 1 with 10 normal distributed classes, model 2 has two normal distributes clusters in every class: a) supervised loss (1), b) unsupervised loss (2), c) validation of model with respect to testing data, d) generator loss (3)

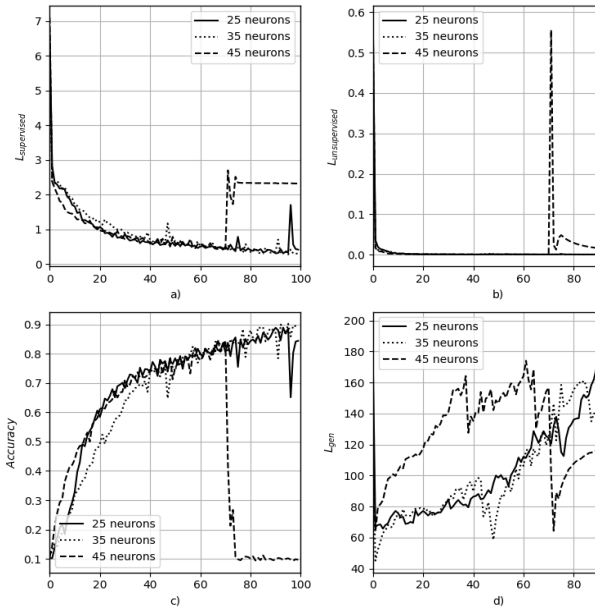

Fig. 2. Visualisation of training process for petrol dataset for different N_{base}:25, 35 and 45 a) supervised loss (1), b) unsupervised loss (2), c) validation of model with respect to testing data, d) generator loss (3)

But for heavy practical real-world examples training GAN-SSL is very difficult. This is due to one big problem for GANs: they convergence very unstable. We can`t get more than 93%.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

The training was performed for a different base number of neurons N_{base}: 25, 35 and 45. The power of the generator and the discriminator must be selected to be the best for the problem of a given dimension, since it can be seen that an overly complex neural network can train unstable. The base neuron parameter for petrol identification problem is selected as 35 neurons.

The total experiments result presents in Table I.

TABLE I. EXPERIMENTAL RESULTS

Base number of neurons	**Accuracy**	**Epochs**
Model 1 dataset		
25	0.999	30
Model 2 dataset		
15	0.977	52
25	0.997	27
Petrol dataset		
25	0.883	200
35	**0.934**	**150**
45	0.906	300

The good situation if we have half of labeled data in whole quantity. But the part of labeled data in data set can be small enough if there are no good labeled objects in consideration.

V. CONCLUSIONS

In this research we investigate the GAN-SSL performance for identification expertise classification problem. This work has showed that GAN-SSL classifiers converge quickly and have good model quality for good normal distributed classes with the same number of examples. The classification model for petrol identification expertise was created and has 93% quality but convergences training was much worse. The number of neurons N_{base} needs to be adjusted.

The identification expertise problem is challenging for classification because of complex structure of classes, outliers and cross-classes. In future work we plan to use the "generated fake" examples to generate missing data to reconstruct certain probability distribution p_{data} of difficult for classification classes.

In this work we use GAN-SSL classification on petrol identification expertise example, but this classification model can be used for diesel fuel, household chemicals items, different oils and for various other objects.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, Sh. Ozair, A. Courville, Y. Bengio, "Generative adversarial nets," Advances in Neural Information Processing Systems 27. Curran Associates, Inc., 2014. pp. 2672–2680.

[2] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, "Improved techniques for training GAN," Advances in Neural Information Processing Systems, 2016, pp. 2234–2242.

[3] A. Maksimova, "Formal statement of the problem of identification expertise, " Donetsk Readings 2017: Russian World as a Civilizational Basis for Scientific, Educational and Cultural Development of Donbass, Donetsk, pp. 69–70.

[4] A. Maksimova, "The approach to the construction of information automated systems of identification examination based on machine learning methods," International scientific and technical congress Intellegent systems and information technologies", Taganrog, 2017, Vol. 1, pp. 438–443.

[5] A. Maksimova, "Fuzzy approach to solve pattern recognition problem for automatisation system for identification expertise (example for petrol identification), " International scientific conference Computer Science and Information Technology, Saratov, 2016, pp. 256–258.

[6] J Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks, " 2016, URL: https://arxiv.org/abs/1511.06390.

[7] D. Kingma, D. Rezende, S. Mohamed, M. Welling, "Semi-Supervised Learning with Deep Generative Models Proceedings of the International Conference on Machine Learning, ", 2014, pp. 3581–3589.

[8] M. Ding, J. Tang, J. Zhang, "Semi-supervised learning on graph with generative adversarial nets, " ACM int. conf. on information and knowledge management, 2018, pp. 913–922.

[9] V. Kovalev, S. Kazlouski, "Examining the capability of GANs to replace real biomedical images in classification models training, "

[10] S. Ioffe, Ch. Shegedy, "Batch normalisation: accelerating deep network training by redusing interval covariance shift," 2015, in ICML`15, pp. 448–456.

[11] T. Saliman, D. Kigma, "Weight normalisation: a simple reparametrization to accelerate training of deep neural networks," in NIPS`16, pp. 901–909.

[12] X. Liu, X. Xiang, "How does GAN-based semi-supervised learning work?", 2020, URL: https://arxiv.org/abs/1809.00130.

[13] D. Kigma, J. Ba, "Adam: a method for stochastic optimization," 2015, Computer science, mathematics, n.pag.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.