






Validity and Limitations of Supervised Learning for Phase Transition Research

Diana Sukhoverkhova^{1,2}, Vladislav Chertenkov^{1,2}, Evgeni Burovski^{1,2},
and Lev Shchur^{1,2}

¹ Landau Institute for Theoretical Physics, Chernogolovka, Russia
lev@landau.ac.ru

² HSE University, Moscow, Russia

Abstract. We analyze the Ising model and the Baxter-Wu model in two dimensions using deep learning networks trained to classify paramagnetic (PM) and ferromagnetic (FM) phases. We use the usual Metropolis Monte Carlo algorithm to create uncorrelated snapshots of spin states. The images used as training data are labeled as belonging to the PM state or the FM state using analytically known phase transition temperatures depending on a given set of parameters. The main result of the paper is that the widely used technique for extraction of the critical temperature directly from the dependence of the output function is not universal. The value of the output function at the critical temperature really depends on the anisotropy of the model under study, the architecture of the deep network, and some parameters of the deep network application.

[AQ1](#)

Keywords: Machine learning · Anisotropic models · Ising model · Baxter-Wu model · Phase transitions · Critical temperature · Critical exponents

1 Introduction

Machine learning has become the *fourth* paradigm of scientific research, following the historically introduced 1) *experiment*, based on practical experience in contact with nature, 2) *theory*, which developed models and mathematical methods to explain some experiments, and 3) *computer simulations*, which is an extended application of models and applied mathematics to situations in which the theory could not make any clear predictions due to mathematical limitations. From about 2015 A.D. It became clear that data-driven science based on deep machine learning is a powerful research tool and can complement the other three paradigms.

It seems relevant to conduct research on the application of new methods of deep machine learning to the problems of natural science in order to verify the accuracy of the extracted results and search for the possibility of obtaining new knowledge. It is also important to conduct a comparative analysis of the amount

of use of computer resources when using already proven methods and algorithms and when using methods based on deep machine learning.

In this paper, we report the results of such an analysis within the framework of the “statistical mechanics and machine learning” direction. We choose two basic models of statistical mechanics for which there is a deep knowledge of the nature of the phase transition and a number of precise results, including a full set of analytical knowledge and reliable computational methods. These are the Ising model with a number of parameters and the Baxter-Wu model. These models belong to different classes of universality. For two of these models, a complete mathematical description of phase transitions is known, including the dependence of the critical temperature on the model parameters and a set of critical exponents that describe the universal behavior of thermodynamic quantities near the critical temperature. We choose three network architectures, CNN, FCNN and ResNet for model analysis.

The paper is organized as follows. Section 2 summarizes previous major work on applying machine learning to statistical mechanics models. The Sect. 3 presents the analytical knowledge that will be used to generate data and analyze the output of the deep network. Section 4.2 describes the deep learning networks used in the analysis. Section 6 presents the results of testing the Ising model taking into account the anisotropy of spin interactions, which leads to a deviation from the critical temperature prediction. Section 7 discusses how the number of epochs affects the shape of the output function. Section 8 summarizes the results obtained and discusses further work.

2 Previous Work

The first paper in the field [1] reports on the application of CNN to classify the paramagnetic (PM) and ferromagnetic (FM) phases of the Ising model. The main results are: 1) the output function is equal to $1/2$ at the critical temperature and 2) the data collapse of the output function obtained for various lattice sizes gives an estimate of the correlation length exponent.

The number of papers follows which use that idea with application to Potts model at second order and first order phase transitions [2]. The determination of the Berezinsky-Kosterlitz-Thouless phase transition and the second-order phase transition in XXZ models was recently reported [7], and the CNN network from the Keras library was used to classify the PM and FM phases to estimate the critical temperature. Therefore, they used the same network architecture and the same analysis as in the article by Carrasquilla and Melko [1].

Various network architectures, training protocols, and deep learning neural networks (NNs) have been used [3–6] to solve multiple physics problems and using supervised or unsupervised learning.

They all use the approach presented in the pioneering article [1], and the main purpose of this article is a thorough and detailed analysis of the applicability of the approach. In our previous article [8], we report a preliminary study of two-dimensional Ising and Baxter-Wu models. We have found that variation of the output function is more informative than the output function itself.

3 Models

We consider two two-dimensional models: the Ising model and the Baxter-Wu model. In addition, we consider two versions of the Ising model given on a square lattice and solved by Onsager [9], and on a triangular lattice and solved by Houtappel [10]. The difference between the Ising model and the Baxter-Wu model, which is defined on a triangular lattice, is that in the Ising model there is a coupling between two spins, while in the Baxter-Wu model there is a coupling between three spins.

3.1 Ising Models on Square and Triangular Lattices

Two-dimensional Ising model on a *square lattice* is defined by the Hamiltonian [9] with spins $\sigma_{x,y}$, which takes two values $\sigma_i = \pm 1$ and placed at the vertices of the lattice shown on the left side of Fig. 1

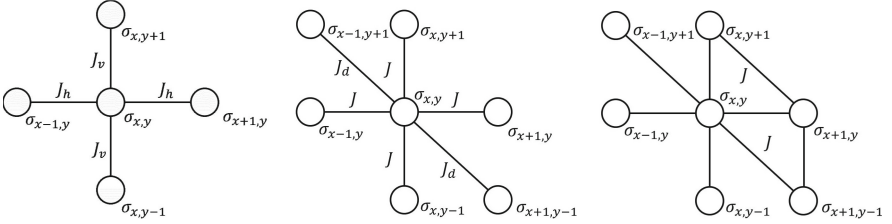


Fig. 1. Illustration of the spin positions and couplings. Left: Ising model on the square lattice, Expr. (1). Center: Ising model on the triangular lattice, Expr. (3). Right: Baxter-Wu model on triangular lattice, Expr. (5).

$$\mathcal{H} = - \sum_{(x,y)} (J_h \sigma_{x,y} \sigma_{x+1,y} + J_v \sigma_{x,y} \sigma_{x,y+1}), \quad (1)$$

where (x,y) denotes the summation over all vertices, J_h is the coupling constant between the horizontal bonds, and J_v is the coupling constant between the vertical bonds of the square lattice. The couplings J_h and J_v are positive, which leads to ferromagnetic ordering of neighboring spins. We use periodic boundaries in both directions.

The critical temperature T_c of the Ising model with Hamiltonian (1) is known [9] from the expression

$$\sinh \frac{2J_h}{k_B T_c} \sinh \frac{2J_v}{k_B T_c} = 1, \quad (2)$$

where T is the temperature and k_B is the Boltzmann factor.

Two-dimensional Ising model on *triangular lattice* defined by the Hamiltonian [10] with the spins $\sigma_{x,y}$, taking two values $\sigma_i = \pm 1$ and placed at the vertices of the lattice shown on the right side of Fig. 1

$$\mathcal{H} = - \sum_{(x,y)} (J\sigma_{x,y}(\sigma_{x+1,y} + \sigma_{x,y+1}) + J_d\sigma_{x,y}\sigma_{x+1,y+1}), \quad (3)$$

where J is the coupling constant between the horizontal and vertical bonds of the square lattice and J_d is the coupling constant between the spins along the diagonals. The J coupling is positive, which leads to ferromagnetic ordering of neighboring spins, while the J_d coupling can be positive, zero, or negative. In the latter case, it will try to make the antiferromagnetic ordering of the spins along the diagonal. For the lack of simplicity, we fix $J = 1$ and vary J_d as a model parameter. We use periodic boundaries in both directions.

The critical temperature of the Ising model with Hamiltonian (3) is known [10] from the expression

$$\left(\sinh \frac{2J}{k_B T_c} \right)^2 + 2 \sinh \frac{2J}{k_B T_c} \sinh \frac{2J_d}{k_B T_c} = 1. \quad (4)$$

In what follows, we will measure the temperature in units of energy (or bond, since spin is a dimensionless quantity) and omit the factor k_B .

3.2 Baxter-Wu Model

Two-dimensional Baxter-Wu model on the *triangular lattice* defined by the Hamiltonian [11] with spins $\sigma_{x,y}$, taking two values $\sigma_i = \pm 1$ and placed at the vertices and with three-spin interactions, as shown on the right side of Fig. 1,

$$\mathcal{H} = - \sum_{(x,y)} J (\sigma_{x,y}\sigma_{x+1,y}\sigma_{x+1,y-1} + \sigma_{x,y}\sigma_{x+1,y}\sigma_{x,y+1}), \quad (5)$$

where the sum in the first term goes over triangles with one orientation, and the sum in the second term goes over triangles with the other orientation, as shown on the right side of Fig. 1. The coupling J is positive, which leads to ferromagnetic spin ordering. We use periodic boundaries in both directions.

The critical temperature of the Baxter-Wu model with the Hamiltonian (5) is known [11]

$$k_B T_c = J \frac{2}{\ln(\sqrt{2} + 1)}. \quad (6)$$

3.3 Phase Transitions and Universality

All three models demonstrate a second-order phase transition - the internal energy E is a continuous function at all temperatures, but the heat capacity C has a singularity at the critical temperature and diverges to infinity on both sides of the critical temperature. The difference between the models is in the

different law of divergence. The specific heat of the Baxter-Wu model diverges according to a power law [11] $C \propto 1/|\tau|^{\alpha_{bw}}$, where $\alpha_{bw} = 2/3$. The specific heats of two Ising models, the square lattice Ising model, Expr. (1), and the square lattice Ising model, Expr. (3), diverge logarithmically [9] $C_{is} \propto \ln |\tau|$, where the reduced temperature $\tau = \frac{T-T_c}{T}$ is the dimensionless distance of temperature T from critical temperature T_c . Therefore, it is assumed that $\alpha = 0$ for the Ising model.

The correlation length ξ [12] between the spins (the distance is measured in lattice units $a = 1$) diverges at the critical point with the exponent ν , and $\xi \propto 1/|\tau|^\nu$, and the value of ν for the two models is different, for the Baxter-Wu model it is equal to $\nu_{bw} = 2/3$. For both Ising models, the divergence is the same and $\nu_{is} = 1$.

Two critical exponents, α and ν , determine the class of universality of models [13]. Both Ising models, Expr. (1) and Expr. (3), belong to the same universality class named after the Ising model. While the Baxter-Wu model, Expr. (5), belongs to the universality class of the four-state Potts model [14]. It should be noted that the thermodynamic quantities of the four-state Potts model states have an additional logarithmic dependence on reduced temperature [16], while the Baxter-Wu model does not [11, 15]. The absence of logarithmic corrections makes the analysis of the Baxter-Wu model more reliable [17].

4 Data Generation and Deep Learning

We generate data using the Monte Carlo Markov Chain (MCMC) approach and the generated data is used for supervised training and testing.

4.1 Data Generation

We use Metropolis algorithm for data generation [18]. Each set is generated with the fixed lattice size L and temperature of thermostat reservoir T . The unit of time for the data generation is 1 MCS (Monte Carlo Step) which is L^2 local Metropolis updates. The correlation time [19] between spin states estimated [8] $t_{corr} = L^{2.15}$. We drop out the first 20 t_{corr} MCS giving system to thermalize at the temperature T , and then save spin distribution each $2 \times t_{corr}$ MCS as a black-white image, associate black with spin pointing up ($\sigma = 1$) or white with spin pointing down ($\sigma = 0$). Thus saved images are not correlated and do not produce any systematic bias to the future research.

4.2 Neural Network Architectures and Output Data

We use three neural network (NN) architectures: fully connected NN (FCNN) architecture, convolutional neural network (CNN) architecture, and ResNet [20] architecture. The details was reported in our previous paper [8].

The NN parameters does depend on the system size of the statistical mechanics model. For example, for investigation of the Ising models defined by Expr. (1) and Expr. (3) we use NN consisting with following layers – Conv2d (N64, K2x2, S1) (see Ref. [21]), MaxPool2d (2x2), ReLU, Linear (64x(L/2-1)x((L/2-1),64), ReLU, Linear (64,1), Sigmoid. The outputs of each layer are shown in the table 1 and the last fully connected layer have one output neuron which used as prediction of the tested snapshot to the ferromagnetic (FM) state.

Table 1. Output of CNN layers used to analyze Ising models, where bs is a batch size.

Layer	Output Shape
Conv2d	[bs, 64, $L - 1$, $L - 1$]
MaxPool2d	[bs, 64, $L/2 - 1$, $L/2 - 1$]
ReLU	[bs, 64, $L/2 - 1$, $L/2 - 1$]
Linear	[bs, 64]
ReLU	[bs, 64]
Linear	[bs, 1]
Sigmoid	[bs, 1]

5 Learning and Testing

We form training datasets of size N_d and test datasets of size N_t . Typical values of N_d and N_t are several hundred, the actual value depends on the task and will be given below. Each set contains data generated with specific values of coupling constants J, J_v, J_h, J_d , snapshot temperatures T , lattice size L , and a class corresponding to temperatures within FM phase (Class = 1) or PM phase (Class = 0) of statistical mechanics models. The Class value is used for supervised learning of the NN.

All samples are randomly divided into batches of four snapshots ($bs = 4$) for each training iteration. The loss function is binary cross entropy (BCE).

$$Q(\hat{f}_i, f_i) = -\frac{1}{N_d} \sum_{i=1}^{N_d} \left[\hat{f}_i \ln f_i + (1 - \hat{f}_i) \ln(1 - f_i) \right], \quad (7)$$

where \hat{f}_i is the correct class, f_i is the NN prediction, $\hat{f}_i \in \{0, 1\}$, $f_i \in [0; 1]$. The Adam algorithm is used for weight optimization [22].

The functions of interest for analysis are the average $F(T; L)$ of the output function $f_i(T; L)$, which is the prediction that sample i with lattice size L generated with temperature T , belongs to FM phase

$$F(T; L) = \frac{1}{N_t} \sum_{i=1}^{N_t} f_i(T; L) \quad (8)$$

and its variation, $V(T; L)$,

$$V(T; L) = \sqrt{\frac{1}{N_t} \sum_{i=1}^{N_t} (f_i(T; L))^2 - \left(\frac{1}{N_t} \sum_{i=1}^{N_t} f_i(T; L) \right)^2}. \quad (9)$$

6 Influence of Anisotropy

In this section, we present an analysis of the output function $F(T; L)$ and variations of the output function $V(T; L)$ for two Ising models with Hamiltonians (1) and (3). We train the NN on the symmetric case of links $J_h = J_v$ for the first model and the zero value of the diagonal link J_d in the second model. Therefore, the training sample in both cases has the symmetry D_4 , the lattice is invariant under rotation through the angle $\pi/2$. The test sets have D_2 symmetry, and the lattice is invariant under rotation through the angle π .

6.1 Ising Model on Square Lattice

Ising model on a square lattice with Hamiltonian (1) solved exactly by Onsager [9]. The critical temperature is given in Onsager's article [9]. The NN was trained on 2048 images of the Ising model with symmetrical coupling constants $J_h = J_v$. Testing was carried out on 512 images of Ising models for different values of the couplings $J_v = J_h$.

The pseudo-critical temperature is obtained in two ways, following the methods proposed in [1] and in [8]. The pseudo-critical temperature depends on the lattice size [24] and converges to the critical temperature T_c in the thermodynamic limit of the infinite size of the system. This fact allows us to estimate the critical temperature from the behavior of the pseudo-critical point, which depends on the size of the system as $\propto 1/L^{1/\nu}$, where ν is equal to the critical length exponent (see discussion in Subsect. 3.3). In addition, this dependence allows us to estimate the exponent of the correlation length ν .

The first [1] method for calculating the pseudo-critical temperature $T^*(L)$ for a fixed lattice size L estimated as the intersection point of the functions $F(T; L)$ and $1 - F(T; L)$, which are the FM and FM phase predictions, respectively. We estimate $T^*(L)$ for each value of the ratio of coupling constants $J_v/J_h = 1, 0.75, 0.5, 0.25, 0.125$, and take limit of the infinite system size using the formula [25]

$$T^* = T^*(L) + \frac{A}{L^b}, \quad (10)$$

where b is an estimate for $1/\nu$, and found visible deviation of the predicted critical temperature from the exactly known one [9] at small values of the ratio J_v/J_h , as shown in the Fig. 2. This deviation can be explained due to the dependence of the correlation length dependence on the ratio J_v/J_h . In the paper [26] the

spin-spin correlation function was analytically calculated for the Ising model on the square lattice in the thermodynamic limit

$$C(\sigma(0)\sigma(R)) = \frac{F_{\pm}}{R^{1/4}} + O(1/R^{5/4}), \quad (11)$$

where R is the distance between any two spins $\sigma(0)$ and $\sigma(R)$, and F_+ and F_- are the amplitudes in the FM and PM phases, respectively. The Fig. 1 of the paper [26] shows dependence of the ratio of the amplitudes F_+/F_- which is similar to those deviation shown in the Fig. 2.

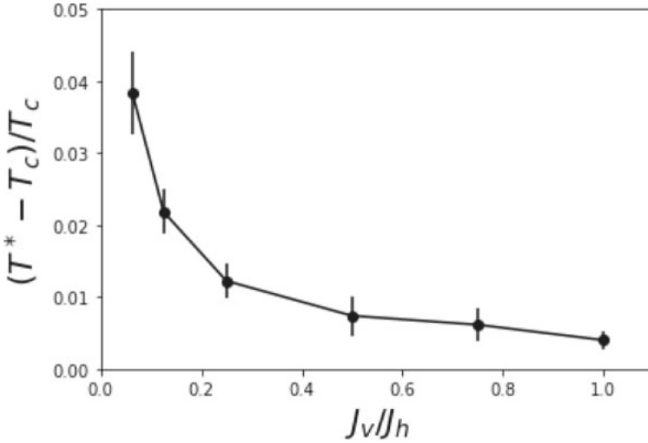


Fig. 2. Deviation of the critical temperature prediction T^* from T_c as function of the coupling ratio J_v/J_h . T^* estimated using method of output functions intersection [1].

The second method for estimating the pseudo-critical temperature [8] is based on the analysis of the variation of the output function $V(T; L)$ (Expr. 9). Approximation of $V(T; L)$ by a unnormalized Gaussian function gives the mean value μ , standard deviation σ (not to be confused with spin) and scale k . The value can be related to the pseudo-critical temperature $T^\oplus(L)$ and an approximation using Expr. (10) gives an estimate of the exponent ν and the critical temperature T_c . The deviation of T^\oplus from the exact T_c is shown in the Fig. 3.

It is noteworthy that the deviations $(T^* - T_c)/T_c$ and $(T^\oplus - T_c)/T_c$ behave qualitatively in the same way, although the second estimation method gives somewhat smaller values.

Table 2 shows estimates of the inverse correlation length exponent $1/\nu$ obtained from the σ variance and demonstrating fairly good agreement with the exact value $1/\nu = 1$ for all ratios of coupling constants J_v/J_h .

6.2 Ising Model on Triangular Lattice

Another interesting and non-trivial model belonging to the universality class of the Ising model, which exhibits the same behavior near the critical point

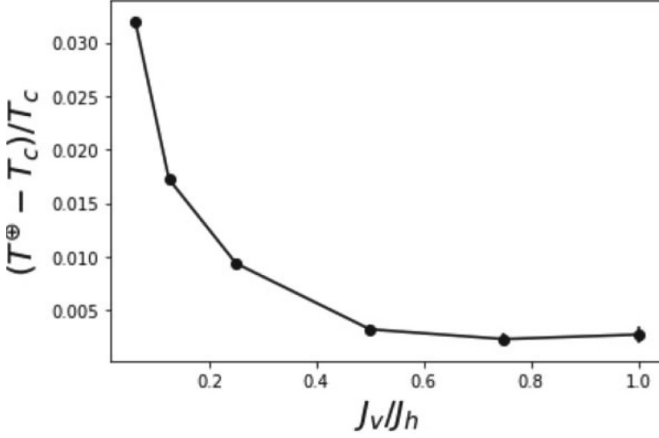


Fig. 3. Deviation of the critical temperature prediction T^\oplus from T_c as function of the coupling ratio J_v/J_h . T^\oplus estimated using approximation of the function $V(T; L)$ [8].

Table 2. The b^\oplus estimate for the inverse correlation length exponent $1/\nu$ obtained from the σ variance.

J_v/J_h	b^\oplus
1.0	1.12(3)
0.75	1.09(3)
0.5	1.07(4)
0.25	1.06(6)
0.125	0.98(14)
0.0625	1.02(9)

in terms of critical exponents, is the Ising model on a triangular lattice with Hamiltonian (3). It was found [27] that the diagonal term, which is proportional to the coupling constant J_d , violates the universality of the Binder cumulant due to significant anisotropy. This case differs from that described in the previous section, for which the Binder cumulant retains a universal value for all ratios of the coupling constants J_v/J_h . The Fig. 4 shows the dependence of the ratio $(T^\wedge - T_c)/T_c$ on the change in the value of the coupling constants J_d/J . The deviation of the predicted critical temperature T^\wedge from the exact one T_c systematically increases with the value of the anisotropy coupling constant.

7 Influence of Number of Epochs for Training

In the previous sections, we have considered the problem of correctly extracting the critical temperature and the correlation length exponent from the output classification function and have analyzed some properties of transfer learning

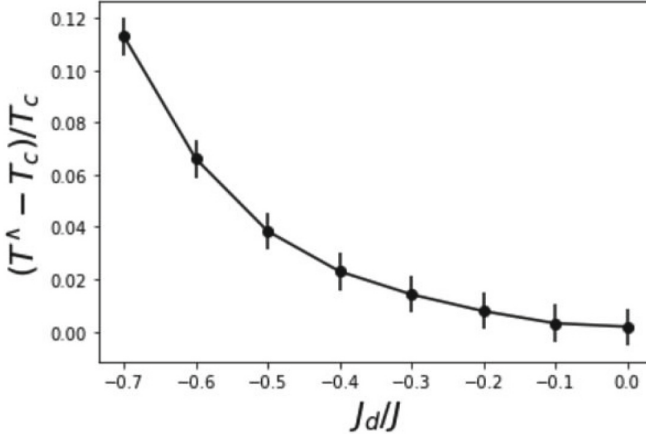


Fig. 4. Deviation of critical temperature prediction T^{\wedge} from exact T_c depending on the coupling ratio J_d/J . Estimation of T^{\wedge} by the intersection of output functions [1].

related to the same problem. The analysis of the previous section is presented in the language of statistical physics. This section is more related to machine learning itself, and primarily deal with the effect of the length of training: the number of training “epochs”.

Transfer learning is a technique commonly used in applications to enable neural networks (NNs) to solve more than one task. This is sometimes referred to as NN pre-training. With pre-training, you don’t have to train a new network for every problem you encounter. Instead, you train NN only once on a wide range of data and then you can use it as a specific layer. This layer, *the backbone*, is a network trained to extract important features. It usually has a deep architecture with millions of parameters and consists of complex layers such as convolutional layer blocks, encoder-decoder blocks, skip connections, attention maps, etc. Often the end layer of a backbone is a vector (embedding vector) that represents various aspects of input object.

The way to learn the critical behavior of spin models with NN classification pre-training translates the problem into multitasking. This study and related works demonstrate the property of the output function that it carries information about the ordered phase, critical temperature, and correlation length exponent. A more detailed study shows that this property is not stable and depends on the quality of NN training.

In previous sections, we analyzed the Ising model. In this section we use the Baxter-Wu model, (5). We expect that the qualitative conclusions for the NN learning process are similar.

7.1 Ordered Phase Prediction in Spin Systems

We have demonstrated that the critical exponent ν can be extracted from a linear approximation on the logarithmic scale of the standard deviation of the $V(T)$

curve, the variance of a ferromagnetic output neuron. It is worth mentioning some aspects of network training that we did not study in the previous sections. The question is how to choose the batch size and how many epochs the training should last.

In [8] we used a batch size of 36. The rule of thumb is that increasing the batch size results in faster learning in terms of CPU time and a sharper decrease in the loss function. The extracted ν values were obtained from the $V(T)$ curve after the first epoch of training, even though the training lasted 10 epochs. Training for more than 1 epoch was necessary in order to make sure that the mean value of the BCE loss function does not grow, and we are not in danger of overfitting.

We challenged the approach used in [8] and ran more experiments with different batch sizes and longer training times in epochs, as shown in Fig. 5. It seemed that 10 epochs was enough to train the network and that the error rate of 0.25 would not drop much in the future. As can be seen from Fig. 5, by the 50th epoch, the error drops 10 times relative to that level to a value of 0.025. A larger batch size results in a faster decrease of the loss function.

This observation raises the question of what would have happened to the functions $F(T)$ and $V(T)$ in epochs 10, 20, 30, 30+ since these functions were used to extract the exponent ν . The Fig. 6 shows the $F(T)$ and $V(T)$ functions predicted on the test data for the Ising model for different epochs with ResNet-10, lattice size 72 and batch size 512. The errors in the figures are less than or equal to the size of the markers.

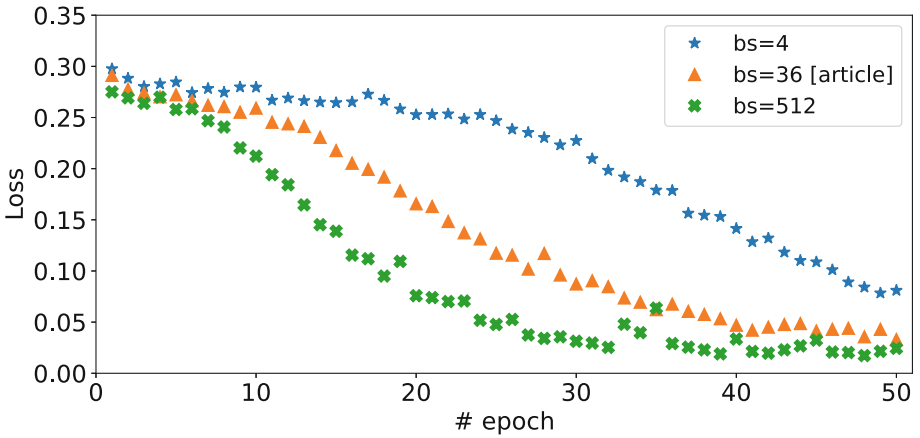


Fig. 5. Validation loss per epoch for different batch sizes (bs) for the Baxter-Wu model, 5.

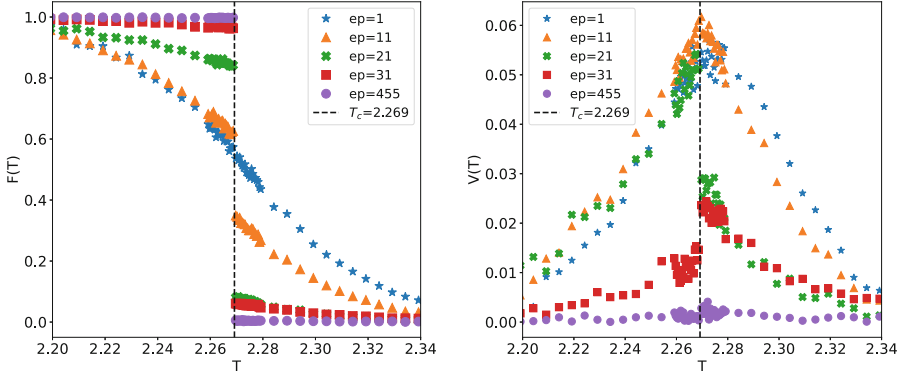


Fig. 6. (left) $F(T)$, the average of the output function, and the variance of output function, $V(T)$, (right) for different numbers of epochs 1, 11, 21, 31, 455. Dashed black vertical line indicates the critical temperature T_c .

The left image in Fig. 6 shows that the output function $F(T)$ turns into a step function as the number of epochs increases. Similarly to $F(T)$, the variance of the output function $V(T)$ in the right image decreases as the number of epochs increases. The direct interpretation is that the NN quality of classification improves. The NN is getting better and better at separating snapshots for the ferromagnetic phase from the paramagnetic ones. The output values f_i become close to 0 or 1. The measure of uncertainty of NN, which is expressed in the function $V(T)$, is reduced. For epochs greater than 1, it becomes difficult and even impossible to analyze the function $V(T)$ using the approximation of the unnormalized Gaussian function mentioned in Sect. 6.1.

This observation about the behavior of the output function $F(T)$ raised another question: what if we train the NN to classify configurations relative to the critical point T_c from the finite-size scaling (FSS) of the thermodynamic quantities. Is the NN able to determine the transition point and whether the step of $F(T)$ is observed near T_c with an increase in the number of epochs?

Let us investigate how the output function $F(T)$ would change from the epoch of the NN and train temperature \hat{T}_c , at which we train NN to classify into ferromagnetic ($T < \hat{T}_c$) and paramagnetic ($T > \hat{T}_c$) phases. Compare the results for different values of \hat{T}_c : a) $\hat{T}_c = 2.269$ from the exact solution, b) $\hat{T}_c = 2.274$ from intersection $F(T)$ with the level 1/2, c) $\hat{T}_c = 2.28$ from the FSS of heat capacity C , and d) $\hat{T}_c = 2.295$ from the FSS of magnetic susceptibility χ . Figure 7 shows the $F(T)$ function predicted on the test data trained at different \hat{T}_c . The errors in the figures are smaller than or equal to the size of the markers.

The NN classifies snapshots with respect to the shifted critical temperature \hat{T}_c . If we select a classification threshold at which all values above are assigned to a positive class, and values below to a negative one, we get the accuracy of correctly classified snapshots close to 100% for epochs greater than 1. The output function $F(T)$ displays the step at the \hat{T}_c used for training for epochs

greater than 1. However, for temperatures $\hat{T}_c = 2.274, 2.28, 2.295$, as the number of epochs increases, $F(T)$ does not turn into a clear step, as at $\hat{T}_c = 2.269$, at which the step at epoch 10 and epoch 30 can be distinguished. At these temperatures, the $F(T)$ function do not differ much between epochs 10 and 80 and does not exhibit a systematic shift as they do at $\hat{T}_c = 2.269$.

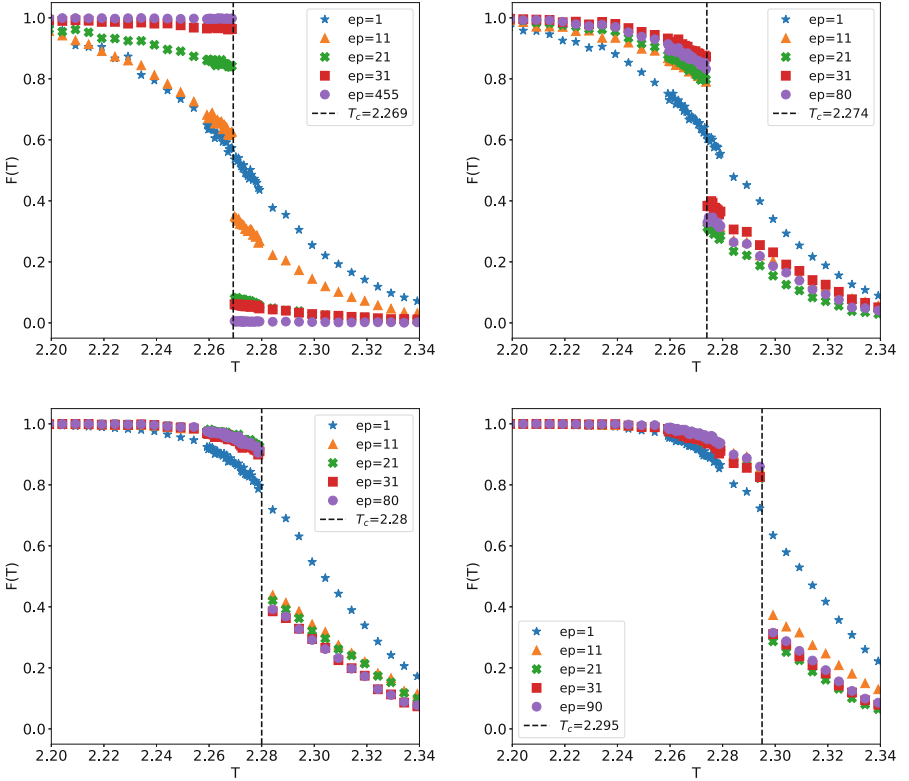


Fig. 7. The output function $F(T)$ for different epochs. Dashed vertical black lines correspond to train temperatures $\hat{T}_c = 2.269, 2.274, 2.280$, and 2.295 in figure order.

8 Discussion

The paper discusses an important issue of applicability of *transfer learning* to critical temperature estimation in statistical mechanics models. Careful analysis using several neural network architectures, several statistical mechanics models, and various methods for extracting the critical temperature estimate from the NN output leads to the following facts.

First, the anisotropy of the interaction of models of statistical mechanics leads to a deviation of the critical temperature estimate from the actual value.

Second, we distinguish two cases of anisotropy – 1) a trivial case of orthogonal anisotropy that can be corrected by an aspect ratio: this case corresponds to the Ising model with Hamiltonian (1); 2) a trivial case of diagonal anisotropy not related to the problem of aspect ratio [27]: this case corresponds to the Ising model with Hamiltonian (3); the critical temperature estimate with the different methods, proposed in [1, 8], coincide well within the statistical and systematic errors of the methods. Third, the critical temperature estimates by different methods proposed in [1, 8] are in good agreement within statistical and systematic errors of the methods. Fourth, the estimate of the correlation length exponent by the method of the article [8] is more reliable than that proposed in the article [1]. Fifth, the estimate of the correlation length exponent for all anisotropic models agrees well with the value of the critical exponent of the correlation length of the Ising universality class, $\nu = 1$.

The batch size and the number of epochs should be chosen to pre-train the classification NN, as they affect the quality of NN predictions. The step function on Fig. 6 is the result of a long training. It should not be treated as overfitting of the model. Most likely, the classification trained NN, loses its generalizing ability, while training, and focuses on optimizing the quality of separating the phases. In Fig. 6, the phases are progressively more accurately separated as the number of epochs increases, and both the NN output, $F(T)$, and its variance, $V(T)$, become unsuitable to extract the critical exponent ν .

Another possible explanation that the output layer of a single neuron is limited by the amount of aspects it contains. An embedding vector of few neurons probably would have greater generalizing ability. A more thorough study is needed, since the huge number of NN parameters makes it impossible to interpret such vectors.

It may be worth focusing on finding more efficient ways to pre-train a NN, in addition to classification. For example, pre-training of language models is often based on understanding the context of a sentence, rather than sentiment classification. The BERT model [23] uses masking of an input for which the network tries to recover a masked part, by itself determining which features should be extracted.

Summing up, there are three main messages of the article. The first positive point is that neural networks trained on an isotropic model predict well the class of universality of anisotropic models. The second negative point is that NN predicts the critical temperature of an anisotropic model with a visible displacement. The third point is that there is some optimum number of epochs for a good estimate of the critical exponent. Therefore, transfer learning is valid for checking the class of universality, and care should be taken if there is no certain knowledge about the anisotropy of the system.

Acknowledgements. Research supported by the grant 22-11-00259 of the Russian Science Foundation.

The simulations were done using the computational resources of HPC facilities at HSE University.

References

1. Carrasquilla, J., Melko, R.G.: Machine learning phases of matter. *Nat. Phys.* **13**(5), 431–434 (2017)
2. Bachtis, D., Aarts, G., Lucini, B.: Mapping distinct phase transitions to a neural network. *Phys. Rev. E* **102**(5), 053306 (2020)
3. Van Nieuwenburg, E.P., Liu, Y.H., Huber, S.D.: Learning phase transitions by confusion. *Nat. Phys.* **13**, 435–439 (2017)
4. Morningstar, A., Melko, R.G.: Deep learning the Ising model near criticality. *J. Mach. Learn. Res.* **18**(163), 1–17 (2018)
5. Westerhout, T., et al.: Generalization properties of neural network approximations to frustrated magnet ground states. *Nat. Commun.* **11**, 1593 (2020)
6. Walker, N., Tam, K.M.: InfoCGAN classification of 2-dimensional square Ising configurations (2020). arXiv preprint [arXiv:2005.01682](https://arxiv.org/abs/2005.01682)
7. Miyajima, Y., Mochizuki, M.: Machine-learning detection of the Berezinskii-Kosterlitz-Thouless transition and the second-order phase transition in the XXZ models. *Phys. Rev. B* **107**, 134420 (2023)
8. Chertentkov, V., Burovski, E., Shchur, L.: Deep machine learning investigation of phase transitions. In: Voevodin, V., Sobolev, S., Yakobovskiy, M., Shagaliev, R. (eds.) *RuSCDays 2022*. LNCS, vol. 13708, pp. 397–408. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-22941-1_29
9. Onsager, L.: Crystal statistics. I. A two-dimensional model with an order-disorder transition. *Phys. Rev.* **65**(3–4), 117–149 (1941)
10. Houtappel, R.M.F.: Order-disorder in hexagonal lattices. *Physica* **16**(5), 425–455 (1950)
11. Baxter, R.J., Wu, F.Y.: Exact solution of an Ising model with three-spin interactions on a triangular lattice. *Phys. Rev. Lett.* **31**, 1294 (1973)
12. Goldenfeld, N.: *Lectures on Phase Transitions and the Renormalization Group*. Addison-Wesley, Reading (1992)
13. Privman, V., Hohenberg, P.C., Aharony, A.: In: Domb, C., Lebowitz, J.L. (eds.) *Phase Transitions and Critical Phenomena*, vol. 14. Academic Press, New York (1991)
14. Potts, R.B.: Some generalized order-disorder transformations. *Proc. Cambridge Philos. Soc.* **48**, 16 (1952)
15. Joyce, G.S.: Analytic properties of the Ising model with triplet interactions on the triangular lattice. *Proc. R. Soc. London A* **343**, 45 (1975)
16. Cardy, J.L., Nauenberg, M., Scalapino, D.J.: Scaling theory of the Potts-model multicritical point. *Phys. Rev. B* **22**, 2560 (1980)
17. Shchur, L.N., Janke, W.: Critical amplitude ratios of the Baxter-Wu model. *Nucl. Phys. B* **840**[FS], 491 (2010)
18. Metropolis, N., et al.: Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087 (1953)
19. Sokal, A.: Monte Carlo methods in statistical mechanics: foundations and new algorithms. In: DeWitt-Morette, C., Cartier, P., Folacci, A. (eds.) *Functional Integration NATO ASI Series*, vol. 361, p. 131. Springer, Boston (1997). https://doi.org/10.1007/978-1-4899-0319-8_6
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)

21. Le Cun, Y., Bottou, L., Bengio, Y.: Reading checks with multilayer graph transformer networks. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, p. 151 (1997)
22. Kingma D., Ba J.: Adam: A Method for Stochastic Optimization, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980)
23. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding (2018) arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805)
24. Fisher, M.E., Ferdinand, A.E.: Interfacial, boundary and size effects at critical points. Phys. Rev. Lett. **19**, 169 (1967)
25. Ferdinand, A.E., Fisher, M.E.: Bounded and inhomogeneous Ising models. I. specific-heat anomaly of a finite lattice. Phys. Rev. B **185**, 832 (1969)
26. Wu, T.T., et al.: Spin-spin correlation functions for the two-dimensional Ising model. Exact theory in the scaling region. Phys. Rev. B **13**, 316 (1976)
27. Selke, W., Shchur, L.N.: Critical Binder cumulant in a two-dimensional anisotropic Ising model with competing interactions. Phys. Rev. E **80**(4), 042104 (2009)