Deep Learning exam, July 2, 2018.

For each lecture there are three questions, and for each paper there is one question, yielding a total of 45 questions. Every question counts equally. There is one correct answer for each question. Write the answer on this paper, in the 'answer' column. Closed book exam: No books, papers, notes, phones, etc allowed. Good luck!

Name: **Student ID:**

Question 1	*Lec 1*	Answer
Why would you use stochastic gradient descent?	**A:** Because it randomizes the computed gradient, avoiding local minima **B:** Because the randomizations makes the resulting approximation function non-linear **C:** Because it reduces the amount of required iterations over the data **D:** Because it reduces the amount of computation per iteration	
Question 2	*Lec 1*	Answer
A linear multi-layered perceptron can be extend to learn nonlinear functions by applying a nonlinear transformation. Which of the following examples does NOT function as a nonlinear transformations?	**A:** Learn the nonlinear function. **B:** Employing a generic kernel function. **C:** Designing feature extractors. **D:** Add recurrent connections.	
Question 3	*Lec 1*	Answer
How can loss be reduced?	**A:** By moving in the opposite direction of the sign of the gradient **B:** By moving in the direction of the sign of the gradient **C:** By finding a global maximum for the loss function **D:** By finding any point where the derivative of the gradient is zero	
Question 4	*Paper 1 Useful Things about ML*	Answer
Having a small dataset of around 200 samples, which evaluation method allows to get most generalizable results?	**A:** Splitting the 80% / 20% of training and validation data and testing on the latter. **B:** Applying 10-fold cross-validation. **C:** Applying Leave-one-out cross-validation. **D:** Training on the entire dataset and investigating the training error (apparent error).	
Question 5	*Paper 2 Deep learning*	Answer
Which of the following functions is NOT commonly used in neural networks to model non-linearities?	**A:** the rectified linear unit ReLU $f(x) = max(0, x)$ **B:** the hyperbolic tangent $f(x) = (exp(x) - exp(-x))/(exp(x) + exp(-x))$ **C:** the sinusoidal function $f(x) = sin(x)$ **D:** the logistic function $f(x) = 1/(1 + exp(-x))$	
Question 6	*Paper 3 Critical appraisal*	Answer
Which one of the following statements is FALSE?	**A:** With millions or even billions of parameters, a neural net is essentially a black box system to humans. **B:** Only when given vast amounts of data, are current deep learning systems able to solve problems concerning open-ended inferences with human-level accuracy. **C:** Deep learning currently has no natural way to deal with hierarchical structures, because all its learned correlations are non-hierarchical. **D:** Currently, deep learning systems learn specific contingencies for particular scenarios, but generally perform poorly when this scenario is changed slightly.	

Question 7	*Lec 2*	Answer
When calculating all derivatives in back-propagation, what trick can be used to make this process more efficient	**A:** Since the chain rule can be used elements that have a negative weight can be skipped, applying this reduces computation time **B:** The weights do not have to be altered during each iteration but only when the error is above a threshold **C:** Since the chain rule is used, calculated elements can be used in determining the next derivative resulting in values only being computed once. **D:** None of the above	
Question 8	*Lec 2*	Answer
Which of the following is the correct formula for gradient descent? (where epsilon is the learning rate and df is the gradient of f)	**A:** theta' = theta - epsilon * df(x) **B:** theta' = theta + epsilon * df(x) **C:** theta' = epsilon * theta - df(x) **D:** theta' = epsilon * theta + df(x)	
Question 9	*Lec 2*	Answer
At what node do we start the backward pass?	**A:** The regularized loss **B:** The input **C:** The class label **D:** The weight	
Question 10	*Paper 4 CNN off the shelf*	Answer
What computer vision problem shows promising results when using off-the-shelf CNN?	**A:** Object detection **B:** Fine grained Recognition **C:** Image classification **D:** All of the above	
Question 11	*Paper 5 Large-scale machine learning*	Answer
What constrains large-scale learning problems?	**A:** The number of examples **B:** The computing time **C:** The implementation of SGD **D:** None of the above	
Question 12	*Paper 6 Vis-Und CNN*	Answer
What is not an application for which the authors used the visualization of activity within a CNN?	**A:** Model architecture selection **B:** Occlusion sensitivity of the model **C:** Analysis of the correspondence specific object parts **D:** Feature generalization	
Question 13	*Paper 7 Yolo*	Answer
Compared to other, state of the art object detectors, what is the main design characteristic of the model setting YOLO apart?	**A:** It divides the image into grid cells for object detection **B:** It is one convolutional neural network that performs all step of object detection **C:** It is a pipeline of several modules, all performing one small task in the object detection task **D:** It outputs a tensor containing info about object location, size, and class	
Question 14	*Lec 3*	Answer
If you convolve an image of 226x226 pixels with a 7x7 kernel and a stride of 2, how large is the resulting convolution?	**A:** 113x113 **B:** 106x106 **C:** 109x109 **D:** 110x110	
Question 15	*Lec 3*	Answer
A CNN uses max pooling of 2x2 and a stride of 3. Assume the input image is padded. By what factor has the size (amount of pixels) changed after the pooling layer?	**A:** 1/4 **B:** 1/6 **C:** 1/9 **D:** 1/3	

Question 16	*Lec 3*	Answer
Convolve filter H over input I and then apply max pooling with width 3. What is the result? (No padding is applied) where I = [200, 2, 199, 300, 50, 10, 70] and H = [1, -1, 1]	**A:** [397, 103, -51, 260, 110] **B:** [397, 260, 260] **C:** [200, 300, 300, 300, 70] **D:** [200, 300, 70]	
Question 17	*Paper 8 IC-STN*	Answer
What is the main difference/advantage of Inverse Compositional Spatial Transformer Networks?	**A:** It learns image warp parameters without warping the image itself **B:** It avoids using an iterative process **C:** The network only makes use of feedforward Learning **D:** It is a direct application of the Lucas Kanade algorithm in a neural network	
Question 18	*Paper 9 Fully convolutional*	Answer
Fully Convolutional Networks use various layers. One of those types of layers is a deconvolutional layer, what is the task of this layer?	**A:** It perfectly cancels the convolutional layer before it, emphasizing the non-linearity in between them. **B:** It is a learned upsampling of the image **C:** It emphasizes a fixed specific location in the image **D:** It convolves with the transpose of the kernel of the convolutional layer before it	
Question 19	*Paper 10 ResNet as Ensemble*	Answer
What is the vanishing gradient problem?	**A:** In some cases, gradient descent can only find a local optimum of the loss function. **B:** In some cases the gradient will point in opposite direction, causing the "learned" value of the weights to vanish. **C:** In some cases, the gradient becomes zero, because of the discontinuity of the loss function. **D:** In some cases, the gradient becomes so small that it barely affects the change in weight value.	
Question 20	*Paper 11 Capsules*	Answer
Which of the following statement is FALSE?	**A:** A capsule is a group of neurons whose activity vector represents the instantiation parameters of a specific type of entity such as an object or an object part. **B:** The length of the activity vector can be used to represent the probability that the entity exists. **C:** The orientation of the activity vector can be used to represent the instantiation parameters. **D:** Active capsules at one level make predictions, via transformation matrices, for the instantiation parameters of higher level capsules. When multiple predictions agree, a lower level capsule becomes active.	
Question 21	*Lec 4*	Answer
In the case of the Exponentially weighted moving average (EWMA), we are given the recursive formula $S_t = (\rho * S_{t-1}) + (1-\rho) * y_t$, where y_t is a value at time t. Given $\rho_1 = 0.5$ and $\rho_2 = 0.96$, which of the two $\rho's$ will produce a smoother curve?	**A:** Both will produce equally smooth curves **B:** $\rho_2 = 0.96$ will produce a smoother curve **C:** $\rho_1 = 0.5$ will produce a smoother curve **D:** It is not possible to infer which ρ will produce smoother curves with the given information	
Question 22	*Lec 4*	Answer
Which of the following effect we get if we use a Stochastic Gradient Descent with Momentum rather than a Stochastic Gradient Descent without Momentum?	**A:** Stochastic Gradient Descent with Momentum generalize better. **B:** Momentum tends to keep traveling in the same Descent direction, preventing oscillations. **C:** Momentum tends to keep traveling in different Descent direction, preventing oscillations. **D:** Momentum is a much faster computation of the gradient.	
Question 23	*Lec 4*	Answer
Question: What is the difference between SGD with momentum and SGD with Adam optimizer.	**A:** Adam uses the second derivative of the gradient. **B:** Adam uses the first derivative of the gradient. **C:** Momentum includes the variance. **D:** Adam includes the second moment.	

Question 24	*Paper 12 Optimization*	Answer
Which of the following is not a reason one would prefer a more advanced gradient descent optimization algorithm?	**A:** dampening oscillations about the optimal descent trajectory; **B:** applying different learning rates for the different parameters; **C:** mitigating the potential aggressive decay of the learning rate; **D:** decreasing the number of optimization parameters.	
Question 25	*Lec 5*	Answer
Which statement is true about regularization in deep neural networks ?	**A:** To avoid local minima, the weights are initialized to zeros in deep neural networks **B:** L2 regularization reduces the norm of the weights exactly to 0, thus avoids overfitting **C:** Dropout can be seen as a process of constructing new inputs by multiplying by noise **D:** Injecting noise in the output of a neural network can also be seen as a form of data augmentation.	
Question 26	*Lec 5*	Answer
How to overcome overfitting?	**A:** By using more data **B:** By reducing the number of features **C:** By reducing flexibility/complexity of the model **D:** All of the above	
Question 27	*Lec 5*	Answer
What is a realistic goal for an effective regularizer?	**A:** Reduce variance and bias significantly. **B:** Reduce variance significantly and bias moderately. **C:** Reduce bias significantly while not overly increasing the variance. **D:** Reduce variance significantly while not overly increasing the bias.	
Question 28	*Lec 6*	Answer
What is one of the flaws of traditional RNNs?	**A:** They cannot remember for too long **B:** There are no flaws: Its Deep Learning! **C:** They require very little data **D:** They're too slow	
Question 29	*Lec 6*	Answer
Which of these statements about RNN's is false?	**A:** RNN's allow the sharing of parameters across different parts of the model **B:** RNN's work very well on sequential data **C:** RNN's do not support the many to one architecture (with more than 1 input layers but only one output layer) **D:** All of the above statements are true	
Question 30	*Lec 6*	Answer
which of the following is a solution to the vanishing/exploding gradient problem of RNNs?	**A:** Use leaky ReLU instead of regular ones **B:** Use skip connections through time **C:** 'Clip' the gradient **D:** All of the above	
Question 31	*Lec 7*	Answer
What is the main principle of a denoising autoencoder?	**A:** After training the network, robustness in increased by adding noise **B:** After the network has generated an output, the noise is optimized to be removed during training **C:** During training, first remove the noise from the input and provide the perfect input to the network **D:** Apply noise to the input and optimize to reconstruct the noise-free input	
Question 32	*Lec 7*	Answer
Which type of autoencoder introduces an explicit regularizer on the code h?	**A:** Undercomplete autoencoder **B:** Denoising autoencoder **C:** Contractive autoencoder **D:** Variational autoencoder	

Question 33	*Lec 7*	Answer
Select the FALSE description of Unsupervised learning?	**A:** Undercomplete hidden layer compress the input **B:** No compression needed for Overcomplete hidden layer **C:** Higher likelihoods give visually better reconstructions than lower likelihoods **D:** Unsupervised features cannot evaluate classification accuracy	
Question 34	*Paper 13 Random weights*	Answer
Which of the following statement is FALSE?	**A:** Certain convolutional pooling architectures can be inherently frequency selective and translation invariant, even with random weights. **B:** Features generated by random weight networks reflect intrinsic properties of their architecture. **C:** Convolution pooling architectures enable even random weight networks to be frequency selective. **D:** Performance of single layer convolutional square pooling networks with random weights are difficult to correlate with the performance of such architectures after pretraining and finetuning.	
Question 35	*Paper 14 Curriculum Dropout*	Answer
How is Curriculum Dropout different from classical Dropout?	**A:** Dynamically increase the number of units that are suppressed as a function of the number of gradient updates. **B:** Dynamically decrease the number of units that are suppressed as a function of the number of gradient updates. **C:** They are the same, however, the supplied data becomes increasingly difficult during the training phase **D:** None of the above	
Question 36	*Paper 15 Structured Receptive Fields*	Answer
What is the main advantage of a receptive field neural network (RFNN), over a regular convolutional neural network (CNN)?	**A:** A RFNN has smaller filters than a CNN, allowing for distinction between finer details of images. **B:** A RFNN have less layers than a CNN for the same performance. **C:** A RFNN performs better than a CNN when first trained with one data set and then with another. **D:** A RFNN outperforms CNNs when the training data is scarce.	
Question 37	*Paper 16 PointNet*	Answer
How does the paper make PointNet invariant to input permutation?	**A:** It sorts the data. **B:** It augments the data with permutations and uses it as a sequence which is then used to train an RNN. **C:** It applies a symmetric function to the data, where the output of the function is invariant to input order. **D:** It augments the data with permutations and trains the network on the extended set.	
Question 38	*Paper 17 Show, Attend and Tell*	Answer
Why are LSTM network suitable as a decoder to produce image captions? It allows the generation of words on every time stamp conditioned on:	**A:** On the particular context vector from the encoder **B:** Previous hidden state of the decoder. **C:** Previously generated words by the decoder **D:** All of the above	
Question 39	*Paper 18 WaveNet*	Answer
Which one of the following arguments does NOT describe a WaveNet?	**A:** It is a deep neural network for generating raw audio **B:** The model is fully deterministic and autoregressive. **C:** It is able to capture the characteristics of different speakers, while allowing to switch between them. **D:** It can also be employed as a discriminative model.	

Question 40	*Paper 19 Vis-Und RNN*	Answer
What are differences between Gated Recurrent Unit (GRU) and Long Short Term Memory (LSTM) networks?	**A:** GRUs have 2 gates instead of 3, one to determine how much previous memory to keep around and another to determine how to combine the new input with the previous memory. GRUs have shown to perform slightly better on smaller datasets. **B:** LSTMs have 3 gates instead of 4, which are regulators of the flow of values that goes through the connections, LSTMs have shown to perform slightly better on larger datasets. **C:** GRUs have a simpler architecture, with the interpretation of computing candidate hidden vector. GRUs have shown to perform slightly better on larger datasets. **D:** LSTMs have a slightly more complex architecture, but generalise better over larger datasets, they have an update and reset gate which determine what information to keep and what to forget.	
Question 41	*Paper 20 Quo Vadis*	Answer
What is the main underlying principle these networks try to achieve?	**A:** Incorporate spatio-temporal information into the predictions of the network **B:** A video is a larger data format than an image and thus requires more parameters to learn **C:** Because videos are 3D the network analyzing it should also have 3D filters **D:** This allows building upon a pretrained CNN without having to learn feature maps from scratch	
Question 42	*Paper 21 Distributed representations*	Answer
Which of the following is not a characteristic of Paragraph vector?	**A:** It is capable of constructing representations of input sequences of variable length. **B:** It is applicable to texts of any lengths: sentences, paragraphs, and documents. **C:** It does not require task-specific tuning of the word weighting function **D:** It may rely on parse trees.	
Question 43	*Paper 22 CycleGan*	Answer
Given the x is the input, y is the ground truth, G(x) is a function which predicts a y given an x and F(y) is a function which predicts an x given a y, which of the following is the forward cycle consistency?	**A:** x –> G(x) –> F(G(x)) –>x **B:** x –> G(x) –> y **C:** y –> F(y) –> x **D:** y –> F(y) –> G(F(y)) –> y	
Question 44	*Paper 23 Split-Brain*	Answer
What modification to a traditional autoencoder architecture is done to obtain a split-brain autoencoder?	**A:** The net is split, resulting in two disjoint sub-networks **B:** The net is split twice, resulting in four disjoint sub-networks **C:** The net is split between random neurons. **D:** The network is split up in one with only biases and one with weights.	
Question 45	*Paper 24 Intriguing Properties*	Answer
What is an adversarial example	**A:** An image perturbed that maximizes the prediction error. **B:** An image perturbed that minimizes the prediction error. **C:** An image with minimum noise in it's label **D:** An image designed to let the network misclassify all images	

End of exam.