NAME OF THE STUDENT	

DEEP LEARNING FOR ARTIFICIAL INTELLIGENCE — Final exam (21/01/2019)

Master's degree in Telecommunications Engineering (MET), ETSETB TelecomBCN, UPC — Coordinator: Xavier Giró

GENERAL INSTRUCTIONS

This exam follows the general ETSETB regulations.

SPECIFIC INSTRUCTIONS

- Write down now your name in the available header of all pages. Unnamed pages will be ignored.
- Adjust to the provided space for your answers. It is enough for the expected level of detail.
- Time available: 100 minutes.
- Maximum grade is 10 points, and each question has a weight of 1 point.
- Answer in the provided sheets of paper, within the available space. Sheets are structured by instructor to facilitate grading.
- You can only use writing material. No other information sources are allowed in this test.
- If you have any question, raise your hand and wait for instructions from the instructor.
- Any attempt of fraud will be persecuted according to the school regulations.

REVIEW & PUBLICATION OF GRADES

- Publication of preliminary grades: Thursday 24th January @ Atenea.
- Access to the corrected exam: Friday 25th January, 6pm @ D5-003.
- Deadline for the submission of appeals and queries: Monday 28th January @ ETSETB Intranet.
- Publication date of reviewed grades: Tuesday 29th January @ Atenea.

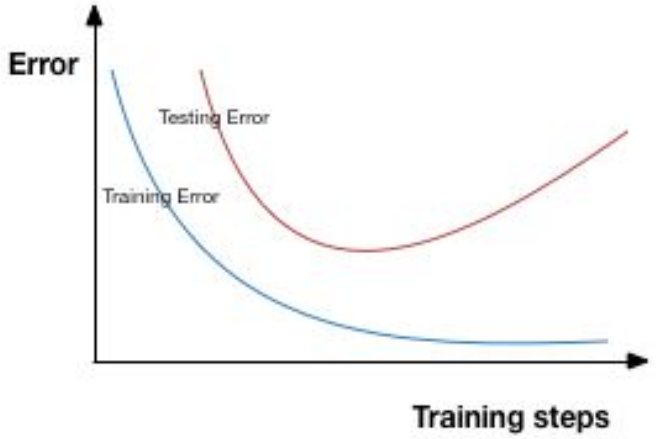
D6L2 INCREMENTAL LEARNING by Ramon Morros

Describe the differences between *multitask learning*, *transfer learning* and *incremental learning*.

NAME OF THE STUDENT	

D7L2 METHODOLOGY by Javier Ruiz

"Early stopping" is one technique to prevent overfitting when training deep networks.The following graph shows the training and validation/testing errors by training steps. Answer the following questions:



a) Explain the "early stopping" mechanism and justify when would you stop if your training has a similar graph as the one above.

b) Enumerate another 3 possible solutions or algorithms to prevent overfitting in your network.

c) Given the possibility, would you use a higher capacity network and prevent overfitting, or a lower capacity one and prevent underfitting?

NAME OF THE STUDENT	

DEEP LEARNING FOR ARTIFICIAL INTELLIGENCE — Final exam (21/01/2019)

Master's degree in Telecommunications Engineering (MET), ETSETB TelecomBCN, UPC — Coordinator: Xavier Giró

D7L1 GENERATIVE MODELS: VAE by Santiago Pascual

A variational autoencoder is a deep generative model with an encoder-decoder architecture, whose loss function is defined as:

$$\log P(X) - D_{KL}[Q(z|X)\|P(z|X)] = E[\log P(X|z)] - D_{KL}[Q(z|X)\|P(z)]$$

(a) What part of the neural network is the term Q(z|X)? and what part is P(X|z)?

(b) Why an autoencoder (NOT variational) is not a generative model?

(c) Why do we normally choose P(z) to be a Normal distribution z ~ N(0, I) ?

D9L1 GENERATIVE MODELS: GANs by Santiago Pascual

In a generative adversarial network (GAN) there are 3 steps within each training iteration (mini-batch) to update both G and D:
(1) D is updated to classify **x** data points from the trainset as real (D(**x**) = 1).
(2) D is updated to classify ẋ = G(**z**) generated data points as fake (D(G(**z**)) = 0).
(3) G is updated to make D(G(**z**)) misclassify its samples as real while D weights are frozen (D(G(**z**)) ~ 1).
In the original "generative adversarial nets" paper by I. Goodfellow et al. they proposed the use of a hyperparameter K during the GAN training that made steps (1) and (2) happen K times prior to going to step (3) in a mini-batch update, as shown in the pseudo-code below:

for number of training iterations **do**

Discriminator training

for k steps **do**

- Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
- Sample minibatch of m examples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ from data generating distribution $p_{\text{data}}(\boldsymbol{x})$.
- Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right) \right].$$

end for

Generator training

- Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
- Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right).$$

end for

Explain why it might be positive for the learning of the generator to update the discriminator K times per each generator update, with K > 1.

NAME OF THE STUDENT	

D9L2 GENERATIVE MODELS: LIKELIHOOD MODELS by Santiago Pascual

WaveNet is an autoregressive deep generative model, which means that for a stream of samples **x** = {x[1], x[2], x[3], …, x[T] } composing the datapoint **x** of T dimensions, each predicted sample depends only on the past (x[t] depends on x[t-1], x[t-2], etc.) . Answer the following questions:

(a) Why does this model need to be autoregressive to represent our data likelihood p(**x**)?

(b) How do we make one-dimensional convolutions causal so that they only depend on past information to predict next sample?

(c) What is a dilated convolution and why is it useful in WaveNet?

(d) What is the main advantage of an explicit likelihood model like WaveNet over GANs?

NAME OF THE STUDENT	

DEEP LEARNING FOR ARTIFICIAL INTELLIGENCE — Final exam (21/01/2019)

Master's degree in Telecommunications Engineering (MET), ETSETB TelecomBCN, UPC — Coordinator: Xavier Giró

D8L1 RECURRENT NEURAL NETWORKS I by Marta R. Costa-jussà

(A) What is one key difference between the backpropagation algorithm and the backpropagation through time algorithm? Briefly justify your answer.

Option 1. Backpropagation through time sums up gradients for corresponding weight for each time step and backpropagation does not do it.

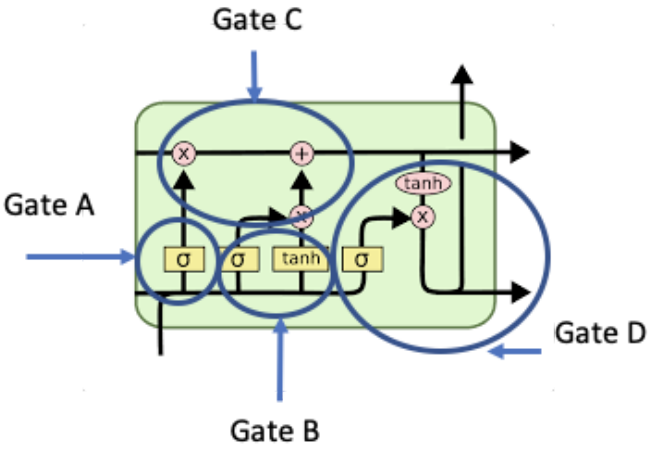
Option 2. Unlike backpropagation, Backpropagation through time does not take into account gradients for corresponding weight for each time step

(B) Why are vanishing gradients a more common problem in basic RNNs compared to feed forward networks?

(C) Mention at least 2 ways of facing the vanishing gradient problem.

D8L2 RECURRENT NEURAL NETWORKS II by Marta R. Costa-jussà

(A) Name each gate in the following scheme of an LSTM:



(B) Explain (with examples) the function of each gate.

NAME OF THE STUDENT	

DEEP LEARNING FOR ARTIFICIAL INTELLIGENCE — Final exam (21/01/2019)

Master's degree in Telecommunications Engineering (MET), ETSETB TelecomBCN, UPC — Coordinator: Xavier Giró

D10L1 ATTENTION-BASED MODELS I by Marta R. Costa-jussà

Decide if the next statements about attention-based models are true or false, and elaborate on your responses.

(A) With an attention mechanism we no longer try encode the full source input into a fixed-length vector. Rather, we allow the decoder to "attend" to different parts of the input at each step of the output generation

(B) Attention has only successfully been applied to text sequences. Other applications such as image or speech have tested attention-based mechanisms but without success.

Correctly Complete the following statements:

(C) The Transformer architecture uses **……..** in the encoder, **……..** attention and **…….** in the decoder

(D) Variations of attention-based mechanisms include: …… attention which is applied to speech recognition and coverage attention which deals with machine translation problems of **…….** and **…….**

D10L2 ATTENTION-BASED MODELS II by Marta R. Costa-jussà

Draw an scheme of how multiplicative attention is computed in an encoder-decoder architecture. Do not forget to mark the key and query vectors. Mention at least a couple of attention score functions.

NAME OF THE STUDENT	

D11L1 REINFORCEMENT LEARNING by Xavier Giró

a) Given a state s, what function specifies which action a to take ?

b) What is the name of the function described in the following expression ?

$$\mathbb{E}_{\pi}[G_t|S_t = s]$$

c) What is the name of the function described in the following expression ?

$$\mathbb{E}_{\pi}[G_t|S_t = s, A_t = a]$$

d) What is the name of the function that quantifies how good or bad an action is with respect to the expectation of returns over all possible actions for a given state ? Provide its expression, too.