TUTOR LESSON 4 - DEEP LEARNING

Exercise 1: How many terms should be summed up to compute $\frac{\partial L}{\partial W}$?
The sequence has 4 terms, minibatch = 1.
(Do not consider the contribution of $h(0)$, which is the zero vector).

SOLUTION:

$$\frac{\partial L}{\partial W} = \frac{\partial L^1}{\partial W} + \frac{\partial L^2}{\partial W} + \frac{\partial L^3}{\partial W} + \frac{\partial L^4}{\partial W}$$

$$= \frac{\partial L^1}{\partial o^1} \cdot \frac{\partial o^1}{\partial W} + \frac{\partial L^2}{\partial o^2} \cdot \frac{\partial o^2}{\partial W} + \frac{\partial L^3}{\partial o^3} \cdot \frac{\partial o^3}{\partial W} + \frac{\partial L^4}{\partial o^4} \cdot \frac{\partial o^4}{\partial W}$$

In general: $$\frac{\partial o^t}{\partial W} = \sum_{t'=1}^{t} \frac{\partial o^t}{\partial h^t} \cdot \frac{\partial h^t}{\partial h^{t'}} \cdot \frac{\partial h^{t'}}{\partial W}, \quad \text{where } \frac{\partial h^t}{\partial h^{t'}} = \prod_{j=t'+1}^{t} \frac{\partial h^j}{\partial h^{j-1}}$$

So: $$\frac{\partial L}{\partial W} = \underbrace{\frac{\partial L^1}{\partial o^1} \cdot \frac{\partial o^1}{\partial h^1} \cdot \frac{\partial h^1}{\partial W}}_{1 \text{ TERM}} + \frac{\partial L^2}{\partial o^2} \cdot \underbrace{\sum_{t'=1}^{2} \frac{\partial o^2}{\partial h^2} \cdot \frac{\partial h^2}{\partial h^{t'}} \cdot \frac{\partial h^{t'}}{\partial W}}_{2 \text{ TERMS}} +$$

$$+ \frac{\partial L^3}{\partial o^3} \cdot \underbrace{\sum_{t'=1}^{3} \frac{\partial o^3}{\partial h^3} \cdot \frac{\partial h^3}{\partial h^{t'}} \cdot \frac{\partial h^{t'}}{\partial W}}_{3 \text{ TERMS}} + \frac{\partial L^4}{\partial o^4} \cdot \underbrace{\sum_{t'=1}^{4} \frac{\partial o^4}{\partial h^4} \cdot \frac{\partial h^4}{\partial h^{t'}} \cdot \frac{\partial h^{t'}}{\partial W}}_{4 \text{ TERMS}}$$

But since $h^0 = 0 \Rightarrow \frac{\partial h^1}{\partial W} = 0$, in fact $h^1 = f(UX_1)$ and does not have dependence on W. This means that:

- For $t=1$: 0 TERMS (no contribution from h^1)
- For $t=2$: 1 TERM (contribution just from h^2)
- For $t=3$: 2 TERMS (contributions from h^2, h^3)
- For $t=4$: 3 TERMS (contributions from h^2, h^3, h^4).

In total: $3+2+1 = 6$ terms.

Exercise 2: Consider a single GRU layer. What are the values to use for z and r to obtain the same behaviour of a standard RNN unit?

a. $z=1$ and $r=1$.
b. $z=1$ and $r=0$.
c. $z=0$ and $r=1$.
d. $z=0$ and $r=0$.
e. It is not possible to reproduce the behaviour of an RNN unit.

SOLUTION:

Remember the equation of a GRU:

$$h^{(t)} = z^{(t)} \odot h^{(t-1)} + (1-z^{(t)}) \odot \underbrace{\tanh\left(Ux^{(t)} + W(r^{(t)} \odot h^{(t-1)}) + b\right)}_{\text{"} \tilde{h}^{(t)}}$$

Remember that the update gate z is making an INTERPOLATION (since we have z and $(1-z)$) between the PREVIOUS HIDDEN STATE $h^{(t-1)}$ and a new CANDIDATE HIDDEN STATE $\tilde{h}^{(t)}$, that may only have the contribution of the input, or also some transformation of the previous hidden state $h^{(t-1)}$, if $r^{(t)} \neq 0$.

The equation of a standard RNN unit is:

$$h^{(t)} = \tanh(Ux^{(t)} + Wh^{(t-1)} + b)$$

$\Rightarrow$ $z=0$ and $r=1$. Answer is (c).

Exercise 3: Consider a CNN layer with 10 filters of size 4X4, a stride of 1 and an input image of size 8X8. How many PARAMETERS are we required to train such a layer? (Do not consider the bias terms).

SOLUTION:

In a CNN layer the LEARNABLE PARAMETERS are:

- SIZE OF EACH KERNEL/FILTER
- # OF FILTERS
- # OF INPUT CHANNELS
- BIAS TERMS (1 for each filter).

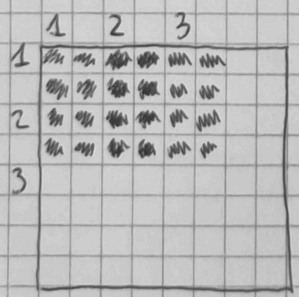
Formula: $(K_w \times K_h \times C_{in} + 1) \times C_{out} = K_w \times K_h \times C_{in} \times C_{out} + \underbrace{C_{out}}_{\text{BIAS TERMS}}$

In our case: $4 \times 4 \times 10 = 160$.

If WE ADD THE BIAS: $160 + 10 = 170$.

Exercise 4: Consider a CNN layer with 10 filters of size 4X4, a stride of 2 and an input image of size 8x8 on a SINGLE CHANNEL. How many MULTIPLICATIONS (between two numbers) are we required to compute the output (feature map) of such a layer? Do not consider bias terms or activation functions.

SOLUTION:



INPUT IMG 8X8

$$\text{Tot.:} \quad \underbrace{4 \times 4}_{\text{FILTER SIZE}} \times \underbrace{3 \times 3}_{\text{HOW MANY TIMES THE FILTER SLIDES IN LENGTH AND WIDTH}} \times \underbrace{10}_{\text{\# FILTERS}} = 1440.$$

- What if we consider the BIAS terms?

The answer DOES NOT CHANGE, because bias terms count as ADDITIONS, not multiplications.

USEFUL FORMULAS :

- CHAIN RULE : $$P(X_1, X_2, \dots, X_m) = \prod_{i=1}^{m} P(X_i \mid X_1, \dots, X_{i-1})$$

with directed (Bayesian models) graphs, this formula becomes:

$$P(X_1, X_2, \dots, X_m) = \prod_{i=1}^{m} P(X_i \mid \underbrace{Pa_G(X_i)}_{\text{PARENTS IN } G \text{ OF } X_i}) \leftarrow \boxed{\text{EACH VARIABLE IS CONDITIONED ONLY ON ITS PARENTS}}$$

- CONDITIONAL INDEPENDENCE : X, Y are INDEPENDENT GIVEN Z $(X \perp\!\!\!\perp Y \mid Z)$ iff

$$P(X, Y \mid Z) = P(X \mid Z) \cdot P(Y \mid Z). \quad (*)$$

- PRODUCT RULE : $$P(X, Y) = P(Y \mid X) \cdot P(X).$$ $$= P(X \mid Y) \cdot P(Y).$$

- BAYES FORMULAS: $$P(X \mid Y) = \frac{P(Y \mid X) \cdot P(X)}{P(Y)} \quad \text{and} \quad P(Y \mid X) = \frac{P(X \mid Y) \cdot P(Y)}{P(X)}.$$

(*) Once you observe the value of Z, knowing the value of Y does not give you any information on X.

Exercise 5: Given stochastic variables X_1, X_2, X_3, X_4 and the fact that

- X_4 is conditionally independent from X_3 given X_2;
- X_3 is conditionally independent from X_2 given X_1;

the joint probability distribution $P(X_1, X_2, X_3, X_4)$ can be factorized as:

a. $P(X_1) \cdot P(X_4 | X_2, X_1) \cdot P(X_2 | X_1) \cdot P(X_3 | X_1)$

b. $P(X_1) \cdot P(X_4 | X_3, X_1) \cdot P(X_3 | X_2) \cdot P(X_2 | X_1)$

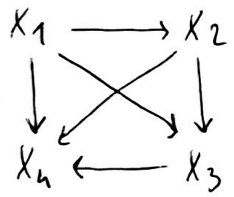
c. $P(X_2 | X_1) \cdot P(X_4 | X_3, X_2) \cdot P(X_3 | X_1) \cdot P(X_1)$

d. $P(X_4 | X_3, X_2)\, P(X_3 | X_2) \cdot P(X_2 | X_1) \cdot P(X_1)$

e. $P(X_3 | X_1)\, P(X_2 | X_1) \cdot P(X_1) \cdot P(X_4 | X_3, X_1)$.

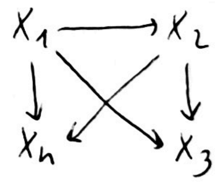
Solution:

- METHOD 1: Starting from the COMPLETE DAG, we remove edges by looking at the information about conditional independence.

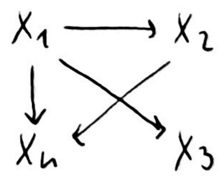


This is the COMPLETE GRAPH. NO CYCLES!
Ordering the variables from the SMALLEST index to the HIGHEST one

① First info: $X_4 \perp\!\!\!\perp X_3 | X_2$, meaning that we can remove the edge $X_4 \leftarrow X_3$ $\Rightarrow$



② Second info: $X_3 \perp\!\!\!\perp X_2 | X_1$, meaning that we can remove the edge $X_3 \leftarrow X_2$ $\Rightarrow$



Now, we can write the joint probability using the formula

$$P(X_1, \dots, X_4) = \prod_{i=1}^{4} P(X_i | Pa_G(X_i))$$

$\boxed{(*) \text{ Note: } X_i \text{ parent of } X_j \Rightarrow i < j}$

$\Rightarrow$ $P(X_1) \cdot P(X_2, X_1) \cdot P(X_3 | X_1) \cdot P(X_4 | X_1, X_2)$.

The correct answer is (a).

• METHOD 2 : Starting from the CHAIN RULE, we rewrite conditional probabilities, by looking at the information.

Start from: $\underbrace{P(X_4|X_3,X_2,X_1)}_{②} \cdot \underbrace{P(X_3|X_2,X_1)}_{①} \cdot \underbrace{P(X_2,X_1) \cdot P(X_1)}_{\text{OK}}$

OK → WE DON'T HAVE INFO about cond. independence between X_1 and X_2

① $P(X_3|X_2,X_1)$

To factorize it, we can use the PRODUCT RULE conditioning on 2 random variables:

$$P(X_3|X_2,X_1) = \frac{P(X_3,X_2,X_1)}{P(X_2,X_1)} \underset{\substack{\uparrow \\ \text{AGAIN} \\ \text{PROD. RULE}}}{=} \frac{P(X_3,X_2|X_1)\cdot \cancel{P(X_1)}}{P(X_2|X_1)\cdot \cancel{P(X_1)}} \quad [P(X_1)>0]$$

$$\underset{[X_3 \perp\!\!\!\perp X_2 | X_1]}{=} \frac{P(X_3|X_1)\cdot \cancel{P(X_2|X_1)}}{\cancel{P(X_2|X_1)}} = P(X_3|X_1) \quad [P(X_2|X_1)>0]$$

$\Rightarrow P(X_3|X_2,X_1) = P(X_3|X_1)$.

② In a Bayesian network, each variable is CONDITIONALLY INDEPENDENT of its NON-DESCENDANTS GIVEN ITS PARENTS.

$$P(X_i \mid X_{j,\, j\neq i}) = P(X_i \mid Pa_g(X_i)) \longleftarrow \boxed{\text{DIRECTLY USE THIS!}}$$

⇒ so conditioning on parents is <u>enough</u>.

Since we know that: $X_4 \perp\!\!\!\perp X_3 | X_2 \Rightarrow X_3 \notin Pa_g(X_4)$ and so

$$P(X_4|X_3,X_2,X_1) = P(X_4|X_2,X_1).$$

This property gives a faster way to solve the exercise: if X_i and X_j are conditionally independent given X_k, then (if $j>i$) X_i <u>cannot</u> be a parent of X_j. This means that in the conditional probabilities of X_j, X_i cannot appear in the conditions.

Exercise 6: Given stochastic variables X_1, X_2, X_3, X_4 and the fact that:

- X_1 is conditionally independent from X_4 given X_2 and X_3;
- X_1 is conditionally independent from X_3 given X_2;

the joint probability distribution $P(X_1,X_2,X_3,X_4)$ can be factorized as:

a. $P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_3,X_2)$

b. $P(X_1)P(X_2)P(X_3)P(X_4|X_3,X_2)$

c. $P(X_1)P(X_2|X_1)P(X_3|X_1)P(X_4|X_1)$

d. $P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_1)$

e. None of the above.

Solution:

Start from $\underbrace{P(X_1)\cdot P(X_2|X_1)}_{OK} \cdot \underbrace{P(X_3|X_2,X_1)}_{①} \cdot \underbrace{P(X_4|X_3,X_2,X_1)}_{②}$

① $P(X_3|X_2,X_1) = P(X_3|X_2)$ because $X_3 \perp\!\!\!\perp X_1 | X_2$

② $P(X_4|X_3,X_2,X_1) = P(X_4|X_3,X_2)$ because $X_4 \perp\!\!\!\perp X_1 | X_2, X_3$.

$\Rightarrow$ $P(X_1)P(X_2|X_1)\cdot P(X_3|X_2)\cdot P(X_4|X_3,X_2)$. Correct answer is (a).

Equivalently: (BY EXCLUSION)

(b) CANNOT be because of $P(X_2)$: we don't have info about conditional independ. of X_1 and X_2, and so for sure $X_1 \in Pa_G(X_2) \Rightarrow P(X_2|X_1)$

(c) CANNOT be because of $P(X_3|X_1)$: we know that $X_1 \perp\!\!\!\perp X_3 | X_2$
$\Rightarrow X_1 \notin Pa_G(X_3)$, so we won't have that term in the factorization.

(d) CANNOT be because of $P(X_4|X_1)$: we know that $X_1 \perp\!\!\!\perp X_4 | X_2, X_3$
$\Rightarrow X_1 \notin Pa_G(X_4)$, so we won't have that term in the factorization.

(e) To exclude this answer, you must be 100% sure that the correct answer is (a) :)