Universidad Técnica Federico Santa María
Departamento de Informática

INF-477 Introducción a las Redes Neuronales Artificiales
Cuestionario II. I-2018.

1. Explique en qué consiste la técnica denominada *Dropout* y qué problema intenta resolver.

2. Suponga que desea aplicar *Dropout* en una determinada capa de una red neuronal convolucional que tiene un gran número de parámetros libres. ¿Deben ser las máscaras estocásticas aplicadas sobre la entrada o la salida de la no-linealidad?

3. Considere una capa convolucional que toma volúmenes de forma $32 \times 32 \times 64$ y produce volúmenes de tamaño 29×29 implementando convoluciones válidas (sin padding). Si se aplica *Dropout* a esta capa, indique qué dimensiones tienen las máscaras estocásticas utilizadas. Indique además, por qué constante deben escalarse los pesos de cada neurona si se emplea un nivel de *Dropout* igual a $p = 0.75$ (probabilidad de retener la neurona).

4. Explique en qué consiste la técnica denominada *Batch Normalization* y qué problema intenta resolver.

5. Suponga que para entrenar una red convolucional muy profunda usted dispone de muy pocos ejemplos etiquetados, pero de muchísimas imágenes no etiquetadas obtenidas del dominio de aplicación. ¿Indique cómo podría utilizar estas imágenes sin etiqueta para mejorar el entrenamiento de su red?

6. Suponga que dispone de una red convolucional muy profunda entrenada durante semanas sobre un conjunto de millones de imágenes de 32×32 pixeles correspondientes a varios tipos de animales, frutas, vehículos, mobiliario doméstico, y otros objetos de la vida diaria. Usted necesita entrenar una red para clasificar imágenes, correspondientes a distintos platos (de comida), que llegan con una resolución de 64×64 pixeles. ¿Tiene sentido utilizar la red ya entrenada para construir su red? Si su respuesta es afirmativa, mencione al menos 3 operaciones que sería necesario llevar a cabo.

7. Suponga que se desea aplicar *weight decay* como método de regularización en una determinada capa de una red neuronal que implementa una transformación de la forma $\sigma(W^T z + b)$. ¿Tiene más sentido aplicar este método separadamente sobre las columnas de W o sobre toda la matriz W simultáneamente? ¿Porqué?

8. Como usted sabe, la función valor absoluto no es diferenciable en 0. En consecuencia, ¿Es correcto matemáticamente hablando utilizar un método como *back-propagation*, basado en gradientes, para entrenar una red neuronal que se está regularizando con la norma ℓ_1? Si es así, ¿Cuál es el costo de esta elección?

9. Explique en qué consiste la técnica denominada *data augmentation* y qué problema intenta abordar. De un ejemplos de operaciones de este tipo que podrían ser aplicadas sobre un dataset visual (imágenes) y sobre un dataset textual (e.g. opiniones).

10. Considere una pequeña red neuronal recurrente con input $x_t \in \mathbb{R}$, capa oculta $z_t = \alpha_1 x_t + \beta_1 z_{t-1} + b_1$ y salida $y_t = \alpha_2 z_t + b_2$. Suponga que después del entrenamiento, los pesos de la red son $\alpha_1 = \beta_1 = \alpha_2 = 1$ y $b_1 = b_2 = 0$. Haga un diagrama de la red indicando claramente los ciclos y pesos correspondientes a estas ecuaciones. Genere además una representación completamente "desenrollada" (unfolded) de la red en el tiempo para la secuencia $x_1 = 2$, $x_2 = -0.5$, $x_3 = 1$, $x_4 = 1$. ¿Qué hace esta pequeña red?

11. ★ Derive explícitamente las ecuaciones correspondientes al backpropagation en el tiempo (BPTT) para una red neuronal diseñada para procesar una secuencia de la forma $x_1, x_2, \ldots x_T$, $x_t \in \mathbb{R}^d$ y cuya

secuencia de salida $y_1, y_2, \ldots y_T$, $y_t \in \mathbb{R}^k$ viene definida mediante las siguientes ecuaciones:

$$y_t = g_o(Vh_t + c) \tag{1}$$
$$h_t = g_h(Wh_{t-1} + Ux_t + b)\,, \tag{2}$$

con $U \in \mathbb{R}^{md}$,$W \in \mathbb{R}^{km}$, $U \in \mathbb{R}^{mk}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^k$.

12. Suponga que se implementa una red neuronal feed-forward para procesar secuencias de largo fijo $T = 10$, con elementos $x_t \in \mathbb{R}^{10}$. Suponga que utiliza una única capa oculta de tamaño 1000 y que la salida es un escalar. ¿Cuántos parámetros más tiene este modelo con respecto a una red recurrente estándar (pregunta 2) con la misma cantidad de neuronas ocultas diseñada para procesar la secuencia? Comente sobre las implicancias computacionales y estadísticas de esta diferencia.

13. Suponga que se desea entrenar una LSTM de 10 celdas que permita anticipar los niveles de una docena de contaminantes atmosféricos en la ciudad de Santiago. Se dispone de datos horarios de esos contaminantes (24 registros por día) para el período que va desde 2006 al 2016. Si representamos esta serie de datos por $(\mathbf{x}_t)_t$, $x_t \in \mathbb{R}^{12}$, indique cómo formaría las secuencias de entrenamiento y qué forma concreta tendría el arreglo de datos que usaría para entrenar el modelo. Indique además los cambios de forma que experimenta el arreglo de entrada hasta que sale de la red.

14. Explique conceptualmente porqué el problema del desvanecimiento o explosión de los gradientes se puede manifestar en redes neuronales recurrentes (RNN), aún si estas tienen una sola capa oculta. Explique además porqué este problema pone límites prácticos al largo de las secuencias que este tipo de redes pueden procesar.

15. Explique porqué la utilización de funciones de activación ReLu se ha mostrado más problemática en redes recurrentes que en redes feed-forward de una profundidad "equivalente". Explique también como podría aliviarse esta dificultad (Hint: considere métodos de inicialización más ciudadosos que los utilizados en redes feed-forward).

16. ★ Explique matemáticamente, porqué el problema del desvanecimiento de los gradientes se puede manifestar en redes neuronales recurrentes, aún si estas tienen una sola capa oculta. (Hint: considerando el modelo de la pregunta 2, derive una cota superior para la magnitud de la matriz $\partial h_t/\partial h_s$, que dependa de $t - s$ y de la magnitud del jacobiano estado-estado $\partial h_{t+1}/\partial h_t$. Para simplificar, puede suponer que función de transferencia aplicada por la capa oculta es lineal. Puede además trabajar con una normal matricial que sea sub-multiplicativa.)

17. En pocas palabras, explique cuál es la "estrategia" implementada por redes ESN (Echo state nets) para permitir el aprendizaje

18. Explique en qué consiste una "unidad permeable" (leaky unit) en el contexto de redes neuronales recurrentes y qué problema se intenta resolver mediante su introducción.

19. Dibuje una típica célula de memoria LSTM que represente claramente el rol de las puertas de entrada, salida y olvido, así como la entrada y salida final del bloque. Escriba además, las ecuaciones que permiten definir formalmente el comportamiento del bloque en el tiempo t, denotando por x_t al patrón de entrada, h_t al patrón de salida y por i_t, o_t, f_t a las activaciones de las compuertas de entrada, salida y olvido respectivamente. Finalmente, explique las ventajas de esta arquitectura con respecto a una red neuronal recurrente clásica.

20. Explique cuál es la diferencia entre una red GRU respecto de una LSTM. ¿Qué objetivo se persigue con esta modificación?

21. Considere una red LSTM con compuertas estándares de entrada, salida y olvido diseñada para procesar una secuencia $x_1, x_2, \ldots x_T$. Construya un diagrama que indique como debiesen organizarse las compuertas de una celda en modo de transmitir intacto un elemento x_{t_1} de la secuencia hasta un tiempo $t_2 \gg t_1$.

22. ★ Considere una red neuronal recurrente de 1 capa oculta, con recurrencias desde la salida, entrenada para producir una secuencia a partir de otra secuencia (many-to-many). Una técnica usada con frecuencia en este caso, consiste en entrenar la red usando *teacher forcing*, es decir, usando la salida correcta en cada tiempo como input para el tiempo sucesivo, en vez de la predicción de la red.

    (a) Explique porqué este criterio es óptimo si se adopta el criterio de estimación de máxima verosimilitud.

    (b) Explique porqué podría ser inconveniente entrenar a la red usando las salidas correctas en vez de sus propias predicciones y proponga un criterio para atenuar este efecto.

23. Suponga que desea entrenar una red LSTM usando *Dropout* sobre los pesos recurrentes. Para una determinada secuencia, ¿Debe ser la máscara correspondiente compartida/fija en el tiempo? o ¿puede aplicarse una realización diferente en cada instante de tiempo?

24. ¿Qué es un modelo neuronal de lenguaje? ¿Porqué se utilizan redes recurrentes para implementar dichos modelos? Cierre su respuesta mencionando una aplicación interesante de los mismos.