UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

Deep Learning

Quiz N°2

Pregrado 2025-o

Apellidos: Méndez Lázaro Nombres: Luis Fernando

Fecha: 21/01/2025

Nota:

Indicaciones:
La Duración es de **30 minutos**.
La evaluación consta de **8 preguntas**.

Las preguntas de opción múltiple valen 1.5 pts.

1. ¿Cuál es la diferencia fundamental entre GRU y LSTM?
   a) GRU elimina la cell state como tal y fusiona su función en un único estado interno simplificado.
   b) GRU conserva tanto la cell state como la hidden state, pero elimina el output gate.
   c) GRU agrega forget gate adicional para el cell state para controlar directamente la salida.
   d) GRU utiliza una cell state interna pero la oculta dentro del reset gate para simplificar la arquitectura.
   e) GRU mantiene la cell state oculta, pero fusiona el input y output gate en una sola compuerta.

2. En un modelo seq2seq con LSTM + Attention, se ha agregado información adicional como features contextuales que no siempre son relevantes para la tarea principal. Sin embargo, el Attention Mechanism puede asignar pesos altos a dichos features irrelevantes. ¿Qué estrategia es más efectiva para forzar la red a discernir la relevancia sin eliminar por completo los nuevos features?
   a) Aplicar un gating adicional (context gate) que combine la hidden state actual con un puntaje de relevancia, el cual será ajustado durante el entrenamiento.
   b) Incrementar el número de capas de LSTM para que la red aprenda a ignorar los features sin utilidad.
   c) Aumentar la diemensionalidad de los features contextuales, de modo que la red identifique más claramente su contenido.
   d) Entrenar una red separada que prediga la relevancia de los features y descartar de forma determinística aquellos con baja puntuación.
   e) Utilizar una máscara binaria diseñada manualmente para anular la atención sobre dichos features irrelevantes

3. En la arquitectura Transformer presentada en "Attention is all you need", consideremos una capa en la que la dimensionalidad de key K, query Q y value V es la misma. Si el resultado de multiplicar QK^T se obtienen valores muy grandes antes de aplicar softmax, ¿qué medida se introdujo para evitar saturación y mejorar la estabilidad del entrenamiento?
   a) Reducción obligatoria de la dimensionalidad de Value para compensar el crecimiento en Key y Query.
   b) Aplicar batch normalization a la matriz de atención resultante.
   c) Establecer una inicialización ortogonal que impida la aparición de valores grandes en las proyecciones.
   d) Un escalado por $d_k^{-1/2}$ que se aplica antes de la operación softmax.
   e) Emplear dropout directamente sobre la matriz Query para reducir la magnitud de las puntuaciones.

4. EfficientNet se basa en un bloque MBConv que incorpora un módulo de Squeeze-and-Excitation (SE). ¿Cuál es su principal?
   a) Reemplazar por completo la convolución depthwise por una operación de gating binaria.
   b) Realizar una reducción dimensional en el espacio de canales mediante convoluciones 1×1 antes de la operación de convolución depthwise.
   c) Implementar una normalización estandarizada por batch para cada canal, mejorando la estabilidad durante el entrenamiento.
   d) Aplicar una atención espacial adaptativa que modula la importancia de diferentes regiones de la entrada.
   e) Reforzar los canales con información más relevante y atenuar aquellos menos informativos.

5. ¿Qué ventaja otorga de MoE y SwiGLU en la arquitectura PaLM?:
   a) Permiten el escalado del número de parámetros sin requerir más recursos computacionales, ya que SwiGLU realiza una compartición de pesos.
   b) Facilitan la activación selectiva de redes expertas y mejoran la expresividad de la capa de activación, al tiempo que reducen el tiempo de inferencia.
   c) Aseguran que todos los tokens pasen por todas las redes expertas, forzando la especialización en cada uno de ellos.
   d) Reemplazan la necesidad de multi-head attention porque cada red experta cumple el rol de una cabeza independiente.
   e) Hacen innecesarias las proyecciones lineales en la etapa de self-attention, ahorrando memoria en el entrenamiento

UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

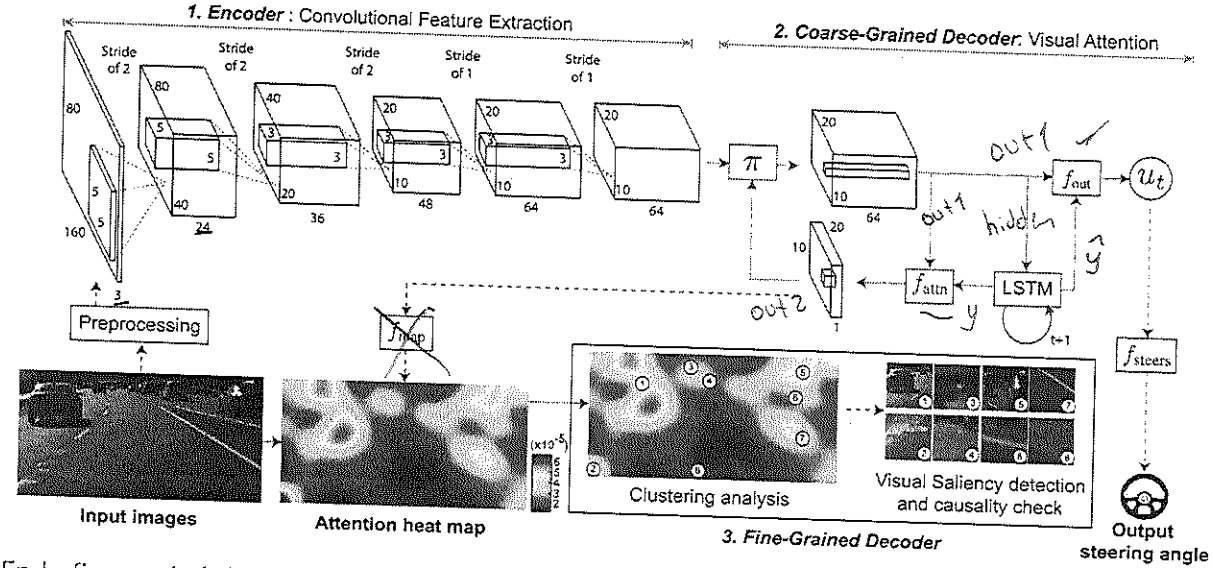
Luis Meléndez

Deep Learning
Quiz N°2
Pregrado 2025-o

6. En BERT, uno de los objetivos de preentrenamiento involucra tareas entre dos oraciones (Next Sentence Prediction). Imagina una base de datos en el que las oraciones no están claramente delimitados. ¿Qué efecto adverso podría surgir en el entrenamiento del modelo?
   a) La red aprende a ignorar el token [CLS], sobrecargando el token [SEP] con toda la señal semántica.
   b) El modelo confunde la relación Next Sentence con relaciones de dependencia de corto plazo, degradando la capacidad de inferencia contextual real.
   c) Los positional embeddings se vuelven irrelevantes al no encontrar límites claros entre oraciones.
   d) La atención se concentra únicamente en las primeras palabras de cada segmento, probocando vanishing gradient en el resto de la secuencia.

7. Acerca de los Forward Problems y los Inverse Problems.
   a) Forward Problems son más fáciles de resolver que los Inverse Problems porque requieren menos capacidad de abstracción.
   b) Forward Problems implican encontrar la causa a partir de los efectos, mientras que los Inverse Problems implican calcular los efectos a partir de la causa.
   c) Los Inverse Problem solo se pueden resolver mediante la intervención humana, ya que requieren creatividad e intuición.
   d) Las IAs son especialmente hábiles para resolver Inverse Problems.
   e) Los Forward Problems y los Inverse Problems son conceptualmente iguales y se pueden resolver con los mismos métodos

8. Para un proyecto de conducción autónoma, se tiene como tarea implementar la arquitectura mostrada en la figura, la cual emplea un mecanismo de spatial attention. Proponga un seudocódigo, asegurando que las dimensiones los tensores internos sean las adecuadas. [10 pts]



En la figura, el símbolo π representa una multiplicación punto a punto en la dimensión espacial. Los bloques f_{attn} y f_{out} son redes que calculan los mapas de atención y calculan la señal u_t, respectivamente. Asuma que la salida de la red es la señal de control u_t. Además, omita el bloque f_{map}.

```
Simple Encoder - Decoder
    init()
        conv1 = Conv2d(3, 24, kernel_size=5, stride=2)
        conv2 = Conv2d(24, 36, kernel_size=5, stride=2)
        conv3 = Conv2d(36, 48, kernel_size=3, stride=2)
        conv4 = Conv2d(48, 64, kernel_size=3, stride=1)
        conv5 = Conv2d(64, 64, kernel_size=3, stride=1)
```

Luis Méndez


```
π = mul()  # multiplicación punto a punto
f_out = Linear( )
f_attn = Conv2d(64, 1, kernel_size=1)   # conv 1x1 Dim reduction!
lstm = LSTM(64)


forward( x )
    x = F.Relu( conv1(x))
    x = F.Relu( conv2(x))
    x = F.Relu( conv3(x))
    x = F.Relu( conv4(x))
    x = F.Relu( conv5(x))

    out1 = π(x)          #branch
    out2 = f_attn(out1)


    y = LSTM(out1)
    y = f_attn(y)
    y = π(y)
return f_out(out1, y)
```