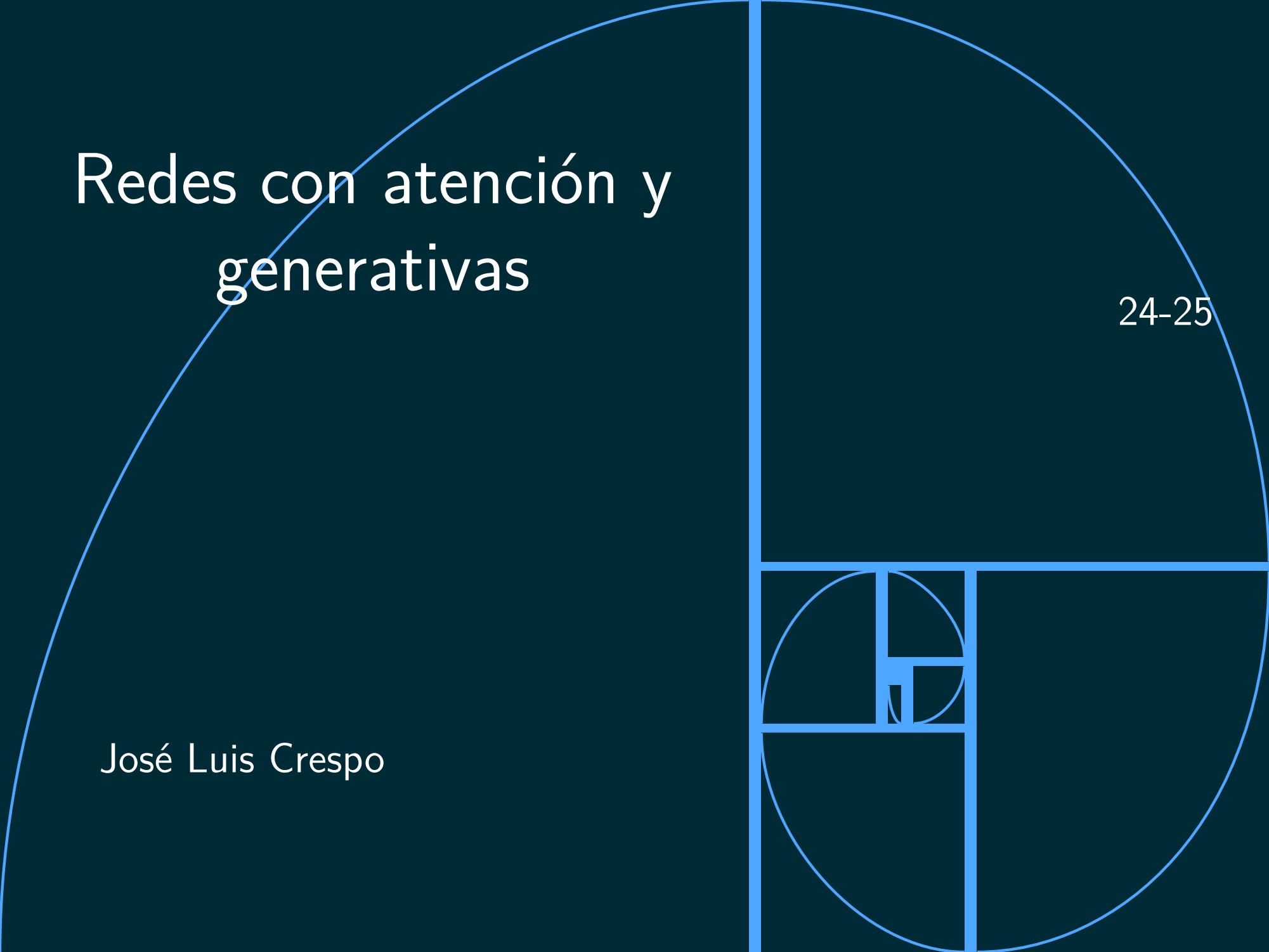
Redes con atención y generativas

24-25



José Luis Crespo

Bloques de atención

Caso normal $Y = f(X)$ siendo f una red

Escoger x mejores $Y = f(X')$ siendo X' una adecuación de X

Bloques de atención

Caso normal $Y = f(X)$ siendo f una red

Escoger x mejores $Y = f(X')$ siendo X' una adecuación de X

Muchas veces $X' = P \cdot X$

Bloques de atención

Caso normal $Y = f(X)$ siendo f una red

Escoger x mejores $Y = f(X')$ siendo X' una adecuación de X

Muchas veces $X' = P \cdot X$

Se puede calcular con unas capas previas

Bloques de atención

Caso normal $Y = f(X)$ siendo f una red

Escoger x mejores $Y = f(X')$ siendo X' una adecuación de X

Muchas veces $X' = P \cdot X$

Se puede calcular con unas capas previas aunque es más complejo

Sistema Q(uery)-K(eys)-V(alues)

Dados q, K, V siendo q una x , K un conjunto de x cuyos valores V tenemos ya

Sistema Q(uery)-K(eys)-V(alues)

Dados q, K, V siendo q una x , K un conjunto de x cuyos valores V tenemos ya

Planteamiento $x' = p \cdot V$ siendo p los pesos de cada uno de las filas de V

Sistema Q(uery)-K(eys)-V(alues)

Dados q, K, V siendo q una x , K un conjunto de x cuyos valores V tenemos ya

Planteamiento $x' = p \cdot V$ siendo p los pesos de cada uno de las filas de V

Pesos $P \propto q \cdot K$

Sistema Q(uery)-K(eys)-V(alues)

Dados q, K, V siendo q una x , K un conjunto de x cuyos valores V tenemos ya

Planteamiento $x' = p \cdot V$ siendo p los pesos de cada uno de las filas de V

Pesos $P \propto q \cdot K$

Lote de muestras $P \propto Q \cdot K, X' = P \cdot V$

Sistema Q(uery)-K(eys)-V(alues)

Dados q, K, V siendo q una x , K un conjunto de x cuyos valores V tenemos ya

Planteamiento $x' = p \cdot V$ siendo p los pesos de cada uno de las filas de V

Pesos $P \propto q \cdot K$

Lote de muestras $P \propto Q \cdot K, X' = P \cdot V$

Caso de usar X $P \propto X \cdot X, X' = P \cdot X$

Capas

- Se alterna un bloque de atención con unas capas convencionales que se aplican sobre cada x' para dar una respuesta

Capas

- Se alterna un bloque de atención con unas capas convencionales que se aplican sobre cada x' para dar una respuesta
- De estos paquetes (atención-capas) se ponen varios

Generación de secuencias

El problema consiste en dada una secuencia, generar otra.

Generación de secuencias

El problema consiste en dada una secuencia, generar otra. Podrían usarse otros tipos de redes, pero vamos a ver las basadas en bloques de atención, que han dado pie a las redes Transformadoras o Generativas

Generación de secuencias

El problema consiste en dada una secuencia, generar otra. Podrían usarse otros tipos de redes, pero vamos a ver las basadas en bloques de atención, que han dado pie a las redes Transformadoras o Generativas Si las secuencias son textos, hay que convertirlas en números para alimentar a la red. Serán vectores de dimensión en los centenares, generados mediante entrenamiento para reproducir las “distancias” (frecuencias de coocurrencia) entre palabras (o fragmentos)

Codificador-Decodificador

- Codificador: a partir de la secuencia de entrada genera una intermedia

Codificador-Decodificador

- Codificador: a partir de la secuencia de entrada genera una intermedia
- Decodificador: a partir de la secuencia intermedia y la parte precedente de la de salida, va generando la de salida

Redes Transformadoras o Generativas

- No hay realimentación, por lo que incluye codificación de la posición (por ejemplo, seno o coseno)

Redes Transformadoras o Generativas

- No hay realimentación, por lo que incluye codificación de la posición (por ejemplo, seno o coseno)
- Codificador: bloque multicapa (atención-red) incluyendo mecanismos para facilitar el ajuste: conexiones puenteadas y normalización

Redes Transformadoras o Generativas

- No hay realimentación, por lo que incluye codificación de la posición (por ejemplo, seno o coseno)
- Codificador: bloque multicapa (atención-red) incluyendo mecanismos para facilitar el ajuste: conexiones puenteadas y normalización
- Decodificador: la secuencia previa pasa por atención y se utiliza como V para otro bloque de atención donde Q, K vienen del codificador; le sigue parte de red convencional. También es multicapa; en todas las capas entra la misma salida del codificador, pero las siguientes capas no toman la secuencia ya pasada, sino la salida de la anterior.