Exposición

0. Introducción (*1 min*)

En este trabajo se estudia la aplicación de técnicas del ámbito del aprendizaje profundo en la generación de música y se desarrolla una herramienta software como ejemplo de aplicación de dichas técnicas.

[PASAR DIAPOSITIVA]

Comenzaremos con algunas definiciones básicas de aprendizaje profundo y a continuación veremos un teorema que caracteriza la capacidad representativa de las redes neuronales, el teorema de aproximación universal. A continuación veremos qué herramientas ofrece el aprendizaje automático en dos campos relevantes para el manejo y la generación de datos musicales. Por un lado, el tratamiento de datos secuenciales, como lo son los musicales. Por otro, el aprendizaje de características para obtener representaciones compactas de datos complejos. Por último estudiaremos el modelo MusicVAE, diseñado para la generación de melodías y veremos su aplicación en la herramienta AutoLoops.

[PASAR DIAPOSITIVA]

1. Aprendizaje profundo (*1 min*)

El aprendizaje profundo es una rama del aprendizaje automático que se basa en al utilización de redes neuronales. El objetivo de estas es aproximar una función de variable real n-dimensional f estrella mediante una función f que depende de unos parámetros, aprendiendo el valor de los parámetros que produzca un mejor resultado.

Se llaman redes ya que suelen representarse componiendo varias transformaciones $f_i : \mathbb{R}^n \to \mathbb{R}^m$, de la forma de las de la diapositiva

$$f_i(\mathbf{x}) = \sigma(\mathbf{W}_i^T \mathbf{x} + \mathbf{b}_i),$$

con $\sigma : \mathbb{R}^m \to \mathbb{R}^m$ una transformación que normalmente es no lineal llamada función de activación. Cada capa de la red representa una de estas transformaciones. $\mathbf{W}$ y $\mathbf{b}$ reciben el nombre de matriz de pesos y vector de sesgos y son los parámetros a optimizar. Esta optimización se realiza mediante versiones del descenso del gradiente, utilizando el algoritmo de propagación hacia atrás para el cómputo del gradiente.

[PASAR DIAPOSITIVA]

2. Aproximación por superposición de funciones sigmoidales (*5 min*)

A continuación veremos un resultado que caracteriza la capacidad de representación que tienen las redes neuronales cuando la función de activación es una función sigmoidal, es decir, una cuyo límite en $-\infty$ es 0 y en $+\infty$ es 1.

2.1. Resultados previos (*1 min 30s*)

Antes de enunciar y demostrar el teorema veamos los resultados previos necesarios.

2.1.1. Teorema de Hahn Banach

El primer resultado es una aplicación directa del teorema de Hahn-Banach, llamada teorema de extensión de Hahn-Banach. Este dice que dado X un un espacio normado, M un subespacio de X y g en el dual de M, existe f en el dual de X tal que f extiende a g, es decir, f y g son iguales en M, y se verifica que $\|f\| = \|g\|$.

[PASAR DIAPOSITIVA]

No aplicaremos este resultado directamente, sino un corolario que nos dice que dado X un espacio normado y M un subespacio de X, si x_0 no pertenece al cierre de M existe un f en el dual de X tal que $f(x) = 0$ para todo x de M, $f(x_0) = 1$ y la norma de f es $\frac{1}{d}$, donde d es la distancia de x_0 a M.

[PASAR DIAPOSITIVA]

Para la demostración de este corolario basta con tomar el conjunto de los puntos de la forma

$$y = x + ax_0$$

y f el funcional que le asigna a a cada punto de esta forma. Se comprueba que f es lineal y que $\|f\| = \frac{1}{d}$. Basta aplicar el teorema de extensión sobre f para obtener el funcional que se buscaba.

[PASAR DIAPOSITIVA]

2.1.2. Teorema de representación de Riesz

El segundo resultado importante es el teorema de representación de Riezs. Dado X un espacio de Hausdorff localmente compacto, y T un funcional lineal acotado en el conjunto de las funciones continuas de soporte compacto sobre X, existe una σ-álgebra en X que contiene todos los conjuntos de Borel de X y existe una

única medida con signo regular μ sobre dicha σ-álgebra de manera que podemos expresar el funcional como la integral de la diapositiva

$$T(f) = \int_X f d\mu.$$

La norma del funcional T es la variación total de μ.

[PASAR DIAPOSITIVA]

2.2. Teorema (*3 min*)

Estos dos serán los componentes principales de la demostración del teorema de aproximación universal.

2.2.1. Función discriminatoria (*30 s*)

Para comprender el teorema debemos definir primero qué es una función discriminatoria. Una función σ de $\mathbb{R}$ en $\mathbb{R}$ es discriminatoria si cuando para una medida μ con signo regular de Borel sonbre I^n se tiene que la integral de sigma del producto estacar de y por x + θ es 0 para todo vector $\mathbf{y}$ y para todo θ, entonces μ es nula.

[PASAR DIAPOSITIVA]

2.2.2. Teorema para funciones discriminatorias (*1 min*)

Podemos enunciar ya el primer teorema. Sea σ una función continua y discriminatoria. Entonces el conjunto de funciones S es denso en el conjunto de funciones continuas sobre I^n, es decir, que dada una f continua en I^n y un ϵ mayor que 0, existe una función g en S tal que el valor absoluto de la diferencia entre $g(\mathbf{x})$ y $f(\mathbf{x})$ es menor que ϵ para todo $\mathbf{x}$ en I^n.

[PASAR DIAPOSITIVA]

Veamos su demostración. S es un subespacio lineal del espacio de funciones continuas sobre I^n. Veamos que si el cierre de S no es igual al conjunto total de las funciones continuas, aplicando el corolario del teorema de extensión de Hahn-Banach obtenemos un funcional lineal acotado no nulo sobre el espacio de funciones continuas en I^n y nulo en el cierre de S. Por el teorema de representación de Riezs podemos expresarlo como la integral de la diapositiva

$$F(h) = \int_{I^n} h(\mathbf{x}) d\mu(\mathbf{x})$$

para alguna medida. En particular, como σ del producto escalar de $\mathbf{y}$ por $\mathbf{x} + \theta$ está en el cierre de S, debe cumplirse que el valor de la integral para esta función es 0, y por tanto ya que σ es discriminatoria, μ es 0. Pero entonces el funcional

F sería nulo, lo cuál es una contradicción. Así, S es denso en las funciones continuas sobre I^n.

[PASAR DIAPOSITIVA]

2.2.3. Lema de funciones sigmoidales (*1 min*)

El siguiente lema nos dice que toda función sigmoidal medible y acotada es discriminatoria, y en particular las funciones sigmoidales continuas lo son.

[PASAR DIAPOSITIVA]

Para su demostración, dada sigma una función sigmoidal, se define la función σ_λ (sigma sub lambda) que converge puntualmente a la función γ (gamma). Comprobemos que σ cumple la definición de función discriminatoria. Tomando pi el subconjunto en el que el producto escalar de $\mathbf{y}$ por $\mathbf{x} + \theta$ es igual que 0 y H en el que es mayor que 0, aplicando el teorema de convergencia dominada obtenemos que la suma de sus medidas es nula. Si esto se cumple para todo $\mathbf{y}$ y para todo θ, la medida debe ser nula y por tanto σ discriminatoria.

[PASAR DIAPOSITIVA]

Definimos el funcional F en las funciones acotadas sobre $\mathbb{R}$, nulo para la función indicadora de cualquier intervalo y por linealidad para cualquier función simple, luego nulo siempre. Tomando las funciones seno y coseno obtenemos que la transformada de Fourier de μ es nula y por tanto μ lo es. Queda probado que σ es discriminatoria.

[PASAR DIAPOSITIVA]

2.2.4. Teorema final y significado (*30 s*)

Basta combinar el primero teorema con el lema recién demostrado para obtener que dada una función sigmoidal continua, el conjunto S es denso en el de las funciones continuas sobre Iˆn. Las funciones de S son aquellas que pueden ser representadas por una red neuronal con una capa oculta de anchura arbitraria. Por tanto este tipo de redes son aproximadores universales.

[PASAR DIAPOSITIVA]

2.3. Otros resultados (*30 s*)

Existen otros resultados sobre la capacidad de representación de las redes neuronales. El recién probado puede generalizarse para funciones acotadas y no constantes o para funciones riemann-integrables no polinomiales. Otro resultado contemplan redes con una anchura de capa fija en función de la entrada y profundidad arbitraria, con funciones de activación ReLU.

[PASAR DIAPOSITIVA]

3. Tratamiento de secuencias (*3 min*)

A continuación vemos el tratamiento de datos secuenciales mediante redes neuronales.

3.1. Redes neurnoales recurrentes

Hasta ahora se han visto redes en las que los datos se tratan de manera aislada. Cada elemento se procesa independientemente sin retroalimentación en la red. Cuando los datos son de tipo secuencial, como por ejemplo textos, audio o música, el contexto de un elemento de la secuencia puede ser importante en su resolución. Además, una secuencia puede tener una longitud indeterminada, por lo que computarlos mediante una red prealimentada profunda puede resultar poco práctico. Para afrontar estas dificultades se introduce retroalimentación en la red, obteniendo redes neuronales recurrentes.

Esta idea es en realidad equivalente a introducir un nuevo parámetro a la función que representa la red, al que llamaremos estado oculto. En cada instante el estado oculto dependerá del estado oculto anterior y la entrada actual. La salida de la red en el instante dependerá de la entrada y del estado oculto anterior.

[PASAR DIAPOSITIVA]

3.2. Redes recurrentes bidireccionales y profundas

Si se quiere capturar información de las entradas pasadas pero también de las futuras para el tratamiento de cada elemento de la secuencia puede utilizarse una estructura bidireccional combinando dos redes recurrentes, una avanzando hacia adelante en el tiempo y otra hacia atrás. La salida de cada instante dependerá de los estados ocultos de ambas subredes.

También es destacable la posiblidad de añadir profundidad a las redes recurrentes. Podemos observar que existen tres bloques funcionales en estas redes: desde la entrada hasta el estado oculto, desde el estado culto previo hasta el siguiente estado oculto y desde el estado oculto hasta la salida. Puede ser beneficioso añadir capas ocultas en cada una de estas estructuras. Sin embargo esto puede dificultar aún más la ya de por sí costosa optimización de las redes recurrentes

[PASAR DIAPOSITIVA]

3.3. Redes recurrentes con puertas

Las redes recurrentes tienen problemas para aprender dependencias a largo plazo ya que por la repetición de una misma operación múltiples veces, el gradiente tiene a desvanecerse o explotar hacia valores muy grandes. Para evitar este fenómeno

se introducen las redes con puertas, que han demostrado ser más efectivas en la práctica. Estas se basan en la idea de crear caminos a lo largo del tiempo a través de los cuales las redes pueden acumular información. Cuándo la información se almacena o deshecha dependerá también de parámetros aprendidos por la red. El modelo más común para redes con puertas es el de LSTM.

El bloque LSTM sustituye a las unidades de la red recurrente. Lo más importante de estas estructuras es que contienen una celda de memoria además del estado oculto. La información que se elimina de dicha celda se decide mediante la puerta de olvido, que depende del estado oculto anterior y la entrada del bloque. A continuación se decide qué información se almacena mediante la puerta de entrada. Por último mediante la puerta de salida se obtiene el nuevo estado oculto, filtrando la información de la celda de memoria utilizando para ello una unidad que depende de la entrada del bloque y del estado oculto anterior.

[PASAR DIAPOSITIVA]

4. Aprendizaje de características (*5 min*)

4.1. Autoencoder (*1 min*)

La efectividad de los métodos de aprendizaje automático depende en gran medida de la representación de los datos que se use. Además, obtener representaciones compactas de datos complejos como la música facilita su manipulación. Es por ello que el aprendizaje de características es un campo de estudio importante. La primera aproximación es la selección de características, que consiste en seleccionar algunas de las ya existentes. Otra opción es generar nuevas características mediante transformaciones de las iniciales.

Existen diversos métodos para generar estas características, como PCA, MDS, Isomap o máquinas de Botlzmann restringidas. En el ámbito del aprendizaje profundo también existe un modelo para este fin, el autoencoder. Consiste en una red neuronal entrenada para copiar la entrada en la salida, obteniendo una representación compacta en el proceso. A la red que lleva la entrada a la codificación se le llama codificador y a la red que lleva la codificación a la salida decodificador.

[PASAR DIAPOSITIVA]

4.2. Autoencoder variacional (*4 min*)

El autoencoder variacional tiene una estructura parecida a la del autoencoder, pero se trata de un modelo generativo. Aprende una distribución de probabilidad de la que generar nuevos datos diferentes a los observados. Asumimos una distribución a priori y una distribución condicional para el decodificador, ambas normales. Querremos obtener la distribución a posteriori para el codificador,

$p_\theta(z|x)$, pero al aplicar la regla de Bayes el cálculo de la distribución marginal del denominador es intratable. Para solucionarlo se introduce un modelo de reconocimiento q, que aproxima la verdadera distribución a posteriori y cuya distribución es normal.

[PASAR DIAPOSITIVA]

Para medir la diferencia entre la aproximación q y la verdadera distribución introducimos una medida de la diferencia entre funciones de distribución, la divergencia de Kullback-Leibler. Para el cálculo de la función de coste del autoencoder variacional nos será útil conocer el valor de la divergencia entre dos distribuciones normales.

[PASAR DIAPOSITIVA]

Obtendremos la expresión de la función de coste minimizando la divergencia entre q y la verdadera distribución p. Sustituímos el segundo término de la esperanza según la regla de Bayes. Extraemos el logaritmo de p de x, puesto que no depende de z, y aplicamos la linealidad de la esperanza. Obtenemos así la función de coste, que recibe el nombre de ELBO. Aplicando el resultado para la divergencia entre normales obtenemos la expresión definitiva de la función de coste a minimizar.

[PASAR DIAPOSITIVA]

La función implica un proceso de muestreo, por lo que no puede calcularse su gradiente. Se sustituye la esperanza por un estimador mediante el truco de reparametrización.

[PASAR DIAPOSITIVA]

5. MusicVAE (*2 min*)

El modelo MusicVAE tiene como objetivo obtener representaciones en características de melodías musicales y generar melodías a partir de estas representaciones. Está basado en un autoencoder variacional, utilizando redes recurrentes en codificador y decodificador.

El codificador utiliza una red LSTM bidireccional de dos capas. El decodificador presenta una modificación en la estructura canónica de la red recurrente para obtener mejores resultados en el muestreo y reconstrucción. El principal objetivo de este cambio es limitar la capacidad de la red LSTM para transmitir información a largo plazo y así forzar el uso de las variables de entrada. Esta modificación se llama decodificador recurrente jerárquico, y trabaja con subsecuencias de la secuencia de entrada. El vector de características z pasa a través de una capa para obtener el estado inicial de la red recurrente directora, que es una red LSTM. Esta produce un vector para cada subsecuencia. Cada uno es tratado individualmente para producir los estados iniciales de una red recurrente que produce la decodificación. Para cada corchea esta capa genera una distribución

en función del vector de la capa directora y de la corchea anterior, de la cual se muestrea la salida.

[PASAR DIAPOSITIVA]

6. AutoLoops (*3 min*)

Explicar el programa