Facultad de Ciencias
Departamento de Física de la Materia Condensada

Trabajo Fin de Grado

DISEÑO BASADO EN REDES NEURONALES DE SENSORES SUPERCONDUCTORES PARA NANOFOTÓNICA Y APLICACIONES CUÁNTICAS

Realizado por:
Marta Sánchez Casi

Dirigido por:
Dr. Sergio Gutiérrez Rodrigo
Dr. Carlos Pobes Aranda

Grado en Física

23 de junio de 2020

Índice

1. Introducción **1**

2. Objetivos **3**

3. Aproximación teórica al problema **5**

3.1. Absorción de luz infrarroja: la nanoestructura 5

3.2. Método FDTD 8

3.3. Redes Neuronales 8

3.4. Optimizador clásico GD 11

4. Resultados **14**

4.1. Obtención de los datos de entrenamiento mediante FDTD 14

4.2. Optimización de la red neuronal 15

4.3. Optimización de los parámetros geométricos del TES 19

5. Conclusiones **23**

1. Introducción

Los sensores de transición superconductora, (TES de sus siglas en inglés *Transition Edge Sensor*), son finas placas de material superconductor que traducen cambios muy pequeños de la temperatura en variaciones medibles de su resistencia eléctrica. Aprovechan el cambio abrupto que se da en la resistencia durante la transición superconductora haciendo del TES un sensor de alta sensiblidad [1].

Los TESs actúan como microcalorímetros con gran resolución en energía, o resolución en número de fotones, y bajo ruido. Un microcalorímetro está compuesto por un termómetro y un material absorbente unido a un baño térmico [2]. De este modo, cuando un fotón llega al dispositivo y este lo absorbe, se produce un cambio de temperatura $\delta T = E/C$, donde E es la energía absorbida y C es la capacidad calorífica del material. Por tanto, tras esta absorción la temperatura total es $T' = T_0 + \delta T$. Para detectar otro fotón de manera aislada, se conecta el dispositivo al baño térmico. Así, el exceso de energía tras absorber el fotón puede fluir hacia el baño térmico, volviendo el sistema a su temperatura original, listo para recibir un nuevo fotón. Sin este baño térmico la temperatura aumentaría en cada interacción con la radiación, operando en este caso el sistema como un detector de integración. En este aspecto, la conductancia G ha de ser elegida de manera adecuada, ya que regula el tiempo de respuesta y el ritmo de detección del sistema.

Se han desarrollado durante los últimos años sensores TES optimizados para la detección de radiación del rango de los rayos X de baja energía [3]. Estos tienen gran interés en astrofísica, donde una parte importante de la información que se recibe del Universo es en esta forma de rayos-X. Se ha propuesto su uso, por ejemplo, en la misión espacial *Athena (Advanced Telescope for High Energy Astrophysics)*, de la Agencia Espacial Europea, cuyo objetivo es el estudio del Universo Energético y Caliente gracias a un telescopio que operará en esta banda de energía [4].

Sin embargo, estos sensores superconductores también pueden ser interesantes en aplicaciones fotónicas, en los espectros ultravioleta (UV), visible (VIS) e infrarrojo (IR) [5] . Un rango muy interesante dentro de estos, es el de las longitudes de onda de las telecomunicaciones, dentro del cual la longitud correspondiente a 1550 nm es clave para varias aplicaciones. Por ejemplo, en esta longitud de onda, las guías de onda, que habitualmente se utilizan en telecomunicaciones, tienen un mínimo de absorción. Por tanto, se puede transmitir información mediante la luz a mayor distancia con menor atenuación. Otra posible aplicación consiste en el intercambio de llave cuántica, que permite intercambiar información de forma segura disponiendo de un canal clásico y uno cuántico, por donde viajarán los fotones. Esto requiere detectar fotones de manera individual con la mayor eficiencia cuántica posible, es decir, es necesario absorber prácticamente el 100 % de los fotones que inciden en el detector.

Los TES se fabrican con distintas combinaciones de metales. En concreto, una de las más prometedoras combina láminas delgadas de molibdeno y oro. Los metales absorben mal en el VIS e IR, muy lejos de las necesidades que las tecnologías discutidas en el párrafo anterior precisarían. Sin embargo, los metales nanoestructurados pueden diseñarse para absorber luz en

esos rangos de energía de forma muy eficiente, gracias sobre todo a la excitación de un tipo de modo electromagnético llamado Plasmón de Supeficie (SPP del inglés *Surface Plasmon Polariton*) [6].

Los plasmones superficiales fueron descubiertos a mitad del siglo pasado por Rufus Ritchie [7]. Según el modelo de Sommerfeld, los metales pueden interpretarse como un gas de Fermi. Los plasmones provienen del acoplo entre los electrones del gas de Fermi y la luz. La característica principal de los SPPs es que son modos evanescentes, lo que implica por un lado que estén muy confinados a la superficie del metal (entre decenas y centenas de nanómetros en el visible), y que por el otro sean fuertemente absorbidos por el metal, dando lugar a longitudes de propagación que van desde las decenas hasta las centenas de micras en el visible [6]. Sin embargo, los SPP's no se excitan simplemente mediante incidencia normal de la luz, ya que este proceso no es capaz de conservar simultáneamente energía y momento. Para que se generen SPPs se deben introducir en la superficie del metal perturbaciones en forma de partículas, agujeros, ranuras... que, gracias a la re-emisión de luz desde esos centros dispersores, permiten el acoplo con el campo evanescente del plasmón de superficie.

2. Objetivos

El objetivo de este TFG es encontrar una propuesta de diseño para un TES sensible a la radiación de longitud de onda de 1550 nm. Los TES que operan en el rango de los rayos X, esquematizado en la figura 1a, cuentan con un absorbente que los hace sensibles precisamente al rango de los rayos X. En este trabajo se busca diseñar un TES sensible al telecom y para ello se propone una nanoestructura metálica, capaz de absorber la luz en forma de SPPs, similar a la que se puede ver en la figura 1b.

Figura 1: Visualización esquemática de la estructura, a grandes rasgos, de dos TES diseñados para distintas aplicaciones: rayos X (izquierda), infrarrojo (derecha). La G es la conductancia del TES al baño térmico a temperatura constante. (a) Esquema de un TES sensible a los rayos X. La radiación incide en el absorbente de rayos X, haciendo el TES sensible a dicha radiación. (b) Esquema de cómo se espera que podría ser un TES sensible a la radiación infrarroja. La radiación incide sobre la nanoestructura metálica de ranuras que permite una alta absorción en este rango.

En el proceso de absorción son clave los SPPs. Se trata de convertir la mayor cantidad de luz en plasmones de superficie. Si se consigue un estado resonante en el que los fotones se acoplen con los plasmones, y estos permanezcan en la superficie el tiempo suficiente, los fotones acabarán siendo absorbidos. De este modo, cuando se excitan los SPPs en metales se tiene una absorción muy alta. Si la superficie metálica del TES no se nanoestructura, la superficie del metal se comporta casi como un reflector perfecto, como se va a mostrar. Por lo que para obtener altas eficiencias de absorción, será necesario modificar la superficie metálica. Por tanto, la finalidad de este trabajo consiste en optimizar las características de estas nanoestructuras para lograr la mayor eficiencia posible. Así pues, una parte importante de este estudio consiste en desarrollar las herramientas numéricas adecuadas para llevar a cabo este proceso de optimización.

Para poder detectar un fotón, además de ser absorbido, este tiene que producir un cambio de temperatura medible. En este aspecto, un TES puede ser sensible a variaciones de hasta unos pocos μK. Si se recuerda la expresión $\Delta T = E/C$, se observa que para pasar de keV (TES sensible a rayos X) a eV (TES sensible al infrarrojo; 0.8 eV para $\lambda = 1550$ nm), es necesaria una disminución de la capacidad calorífica C de los dispositivos de varios órdenes de magnitud,

lo cual impone restricciones sobre su diseño. Una forma de reducir un factor importante la C es eliminando el absorbente, que para el caso del infrarrojo no va a ser necesario. Este factor esencial en el diseño del TES será también tenido en cuenta.

3. Aproximación teórica al problema

El diseño para un TES nanofotónico parte del mismo sistema que se utiliza para detección de rayos X (figura 1a), formado por una lámina de oro sobre otra de molibdeno, con la salvedad de que la capa de absorbente necesaria para las aplicaciones en altas energías no es necesaria en el rango del visible. Las láminas son sub-longitud de onda en el rango de energías de interés, el telecom. La propuesta de este trabajo para conseguir niveles óptimos de absorción en este rango consiste en utilizar una red periódica de ranuras hechas en la capa de oro, capaces de excitar SPPs de forma muy eficiente, tal y como se demuestra más adelante.

Para buscar el diseño apropiado existen programas como el FDTD (*Finite-Difference Time-Domain*) [8] que, dada una nanoestructura como *input*, devuelve como *output* la respuesta óptica del sistema. Sin embargo, el espacio de parámetros a investigar en el proceso de optimización es muy grande (4 parámetros geométricos) y la velocidad de cálculo de estos sistemas mediante FDTD es relativamente lenta (5-6 horas por simulación), por lo que se hace imprescindible una aproximación al problema no basada en "fuerza bruta". No resulta eficiente tratar de encontrar un conjunto de parámetros óptimos mediante prueba-error. De este modo, el proceso de optimización de la nanoestructura se realiza en dos pasos, que en el trabajo se han realizado siguiendo estas técnicas:

1. Redes neuronales: se diseña una red neuronal que, como resultado, devuelve el espectro de absorción del sistema [9]. De esa forma es posible sustituir en los cálculos del espectro de absorción el método numérico FDTD por la red neuronal entrenada, con un incremento muy notable de la velocidad de cálculo del orden de 10^6.

2. Optimizador clásico GD: se utiliza el algoritmo de optimización de descenso de gradiente, (GD del inglés *Gradient Descent*), que emplea la red neuronal para calcular los espectros de absorción de manera rápida.

3.1. Absorción de luz infrarroja: la nanoestructura

Para empezar, antes de introducirse en el funcionamiento y utilidad de los algoritmos mencionados, conviene estudiar en detalle la estructura en la se va a centrar todo el trabajo. Para conseguir la capacidad calorífica C adecuada y la sensibilidad óptima, los TES trabajan a temperaturas muy bajas. Normalmente, no es posible encontrar materiales superconductores con la temperatura de operación requerida , por lo que se emplea lo que se conoce como efecto proximidad [10] en que un metal normal se deposita sobre un superconductor reduciendo su T_c. Así, la T_c final de la bicapa depende de la proporción relativa de espesores.

En el grupo de *Quantum Materials and Devices* (Q-MAD) [11], en el cual se ha realizado este trabajo, lleva desarrollando durante los últimos años sensores TES optimizados para la detección de rayos X de baja energía [3]. Para ello el termómetro se fabrica con bicapas de Au-Mo. Por tanto, la nanoestructura de este trabajo parte de una lámina de oro situada sobre un substrato de molibdeno. Como ya se ha mencionado con anterioridad, a la lámina de Au se le practican unas ranuras de manera periódica. Esto añade otros dos parámetros al problema, además de los espesores de Au y Mo (h_{Au}, h_{Mo}), el periodo p de estas ranuras y su anchura a.

Si el oro no se perforara, la bicapa Au/Mo se comportaría como un espejo casi perfecto, como demuestra la figura 2c, en la que se muestra la absorción en una capa de 100nm de Au y Mo. Así, la nanoestructura se crea mediante la perforación del Au con ranuras, distribuidas periódicamente, como muestra la figura 2, donde también se incluye la notación para los parámetros geométricos. En la misma figura se incluyen también las constantes dieléctricas experimentales para el Au y Mo en el rango UV-VIS-IR (figuras 2a y 2b) [12, 13].

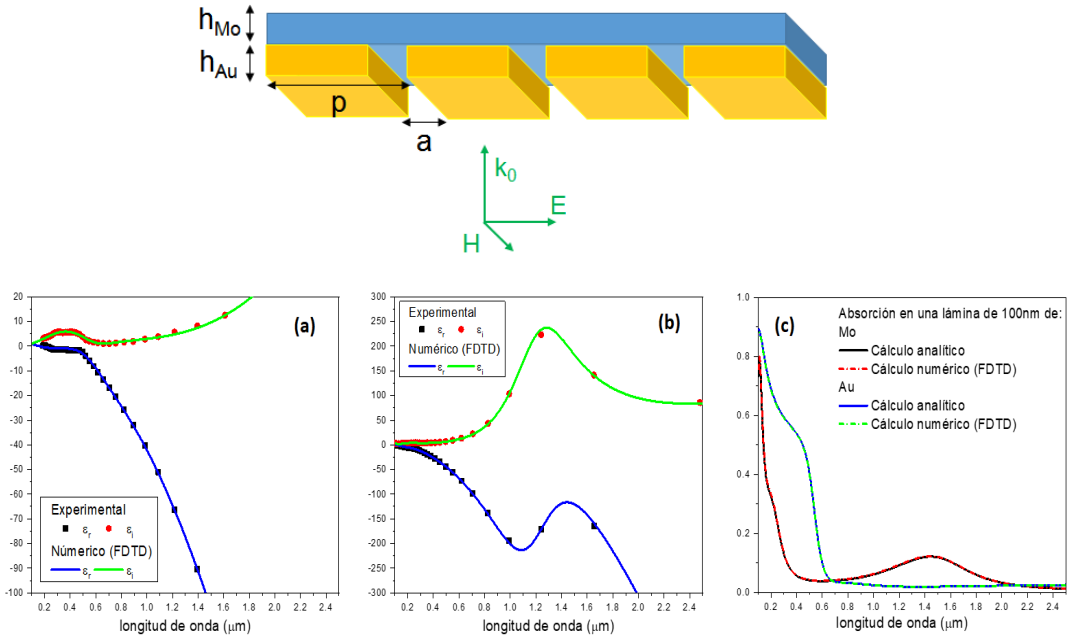

Figura 2: Nanoestructura capaz de soportar la excitación de plasmones de superficie. Consiste en una capa de oro de espesor h_{Au}, perforada periódicamente con orificios de anchura a y periodo p, sobre una capa de molibdeno de espesor h_{Mo}. (a) Constante dieléctrica del Au, partes real e imaginaria, frente a la longitud de onda. Aparece la comparativa entre el cálculo experimental y el obtenido con el algoritmo FDTD. (b) Constante dieléctrica del Mo, partes real e imaginaria, frente a la longitud de onda. De nuevo, se compara el resultado experimental con el resultado del algoritmo FDTD. (c) Curvas de absorción para una capa de 100 nm de Au y Mo. Se observa un muy buen acuerdo entre el cálculo analítico y el obtenido mediante FDTD.

Ya se ha comentado que los TES implementados por el grupo Q-MAD se basan en bicapas de Au-Mo, optimizadas para una temperatura de funcionamiento de 100 mK. Se han conseguido dos combinaciones de espesores de estas capas que trabajen a esa temperatura, y son las siguientes:

- $h_{Mo} = 45\ nm$ y $h_{Au} = 265\ nm$
- $h_{Mo} = 55\ nm$ y $h_{Au} = 340\ nm$

La temperatura de operación para el rango infrarrojo no ha de ser necesariamente la misma, pero se emplean estos valores como una referencia o guía para restringir el espacio de parámetros en el proceso de optimización.

En cuanto a las dimensiones del dispositivo, tanto los límites de fabricación, como la necesidad de incluir un número mínimo de periodos en el metal para obtener una red de ranuras, fijan un tamaño mínimo de dispositivo de unos 10 μm. Mientras que, por otro lado, la necesidad de mantener la capacidad calorífica en valores cercanos a los 10fJ/K impone un tamaño máximo para el TES de unos 100 μm. Por tanto, queda acotado el tamaño del dispositivo que se busca.

En cuanto a las ranuras, se pueden fabricar desde unos cuantos nanómetros hasta unas pocas micras, por lo que es posible realizar ranuras de tamaños del orden de magnitud de la longitud de onda de interés. En este aspecto, como ya se ha comentado, han de caber un número razonable de estas en el metal para que pueda considerarse una red ranuras, normalmente entre 10 y 20 es suficiente.

El fenómeno óptico que opera tras este diseño se conoce como EOT (*Extraordinary Optical Transmission*). En 1944, Hans Bethe, descubrió que la transmisión normalizada al área a través de un agujero circular en una placa conductora delgada es [14]:

$$T \approx \frac{64}{27\pi^2}\frac{r}{\lambda}^4 \tag{3.1}$$

De modo que, siendo r el radio del agujero y λ la longitud de onda de la radiación incidente, si $r << \lambda$ la transmisión es muy débil. Sin embargo, en 1998 Ebbesen y sus compañeros de trabajo descubrieron que era posible la transmisión de luz en el visible e infrarrojo a través de agujeros sub-longitud de onda [15]. Este fenómeno está asociado a una transmisión de la luz a través de agujeros sub-longitud de onda mayor de la que cabría esperar para aperturas de ese tamaño [16]. En principio, como ya se ha visto, esto no se debería dar en agujeros tan pequeños, pero el acoplamiento entre la radiación incidente y los SPPs da lugar a la resonancia necesaria para que el proceso EOT ocurra. Aquí reside la importancia de los plasmones superficiales.

Los picos de EOT están relacionados con la periodicidad de la muestra. Eso es fácil de entender y permite dar una primera estimación de un valor aproximado para el periodo. Cuando la luz llega a la red de ranuras en incidencia normal, para que se exciten los plasmones, se ha de conservar el momento, es decir, el momento de la luz dispersada por el agujero k_s debe ser igual al momento del plasmón k_{SPP}. Al estar en una red de ranuras, en incidencia normal, el momento de la luz dispersada ha de ser un múltiplo entero de $\frac{2\pi}{p}$, de lo contrario, la relación de dispersión se anula y no hay luz dispersada. Por tanto, el momento de los plasmones ha de cumplir $k_{SPP} = n\frac{2\pi}{p}$. Si se considera que la relación de dispersión de los plasmones en superficie plana apenas cambia con la presencia de los orificios y que k_{SPP} se encuentra próximo al cono de luz, se puede sustituir este momento por el de la relación de dispersión de la luz en vacío $\omega = ck_s$. De este modo, tomando n=1, queda lo siguiente.

$$k_{SPP} \approx \frac{\omega}{c} = \frac{2\pi}{\lambda} \Rightarrow \frac{2\pi}{\lambda} \approx \frac{2\pi}{p} \Rightarrow \lambda \approx p \tag{3.2}$$

Así, se obtiene que el valor del periodo ha de ser próximo al de la longitud de onda de resonan-

cia, es decir, $p \approx 1550nm$. Sin embargo, hay que tener en cuenta que esto se ha calculado bajo ciertas aproximaciones, considerando que la relación de dispersión de un plasmón en superficie plana apenas se ve perturbada por la presencia de orificios y que es cercana al cono de luz. Por lo tanto, esto no da el valor exacto del periodo, pero resulta útil para conocer la zona en la que se encuentran las resonancias.

3.2. Método FDTD

El método de diferencias finitas en el dominio del tiempo, (método FDTD de sus siglas en inglés *Finite-Difference Time-Domain*), es uno de los más empleados en electromagnetismo computacional [17]. El algoritmo FDTD es capaz de resolver las ecuaciones de Maxwell numéricamente y, por tanto, proporcionar toda la respuesta óptica de un sistema, incluida la evolución temporal del campo electromagnético. Este método tiene la capacidad de tratar distintos materiales, desde dieléctricos a metales, y diferentes tipos de fuentes. A través de este algoritmo se pueden obtener los coeficientes de transmisión y reflexión del sistema, lo cual permite conocer cuál es la absorción en 1550 nm e ir variando los parámetros del sistema para maximizarla.

Adaptando el algoritmo al sistema de red de ranuras explicado anteriormente, se introducen como *input* los valores del periodo p, anchura del *slit* a y espesores de las capas de oro y molibdeno, h_{Au} h_{Mo}. De este modo, se pueden probar distintas combinaciones de valores de estos parámetros, observar la absorción resultante que ofrece el programa para dichos parámetros e, intuitivamente, variar estos buscando el objetivo.

No obstante, como ya se ha comentado, este método presenta un inconveniente, tarda bastante en resolver el espectro. Además, conforme se aumentan las dimensiones del sistema, este tiempo se dispara llegando a 5-6 horas.

En definitiva, este algoritmo resulta excesivamente lento en procesos de optimización que, como se va a comprobar, requieren miles de cálculos para converger. Por otro lado, la búsqueda intuitiva también es poco eficaz. Por ello, se recurre al uso de las redes neuronales las cuales, una vez entrenadas para calcular el espectro de absorción para un conjunto de parámetros geométricos dados, permiten superar estos inconvenientes.

3.3. Redes Neuronales

La inteligencia artificial se puede definir como el esfuerzo por automatizar, mediante tecnología de computadores, tareas intelectuales normalmente realizadas por los seres humanos [18]. Un tipo de inteligencia artificial lo constituyen las redes neuronales. El término red neuronal hace referencia a la neurobiología, ya que consisten en capas de lo que se llaman neuronas, que se transmiten información entre ellas, como sucedería en un cerebro. Esta idea se ilustra en la figura 3a. Sin embargo, hay que tener claro que las redes neuronales no son modelos cerebrales, lo que se realiza en una red neuronal no es lo que sucede en un cerebro durante el aprendizaje.

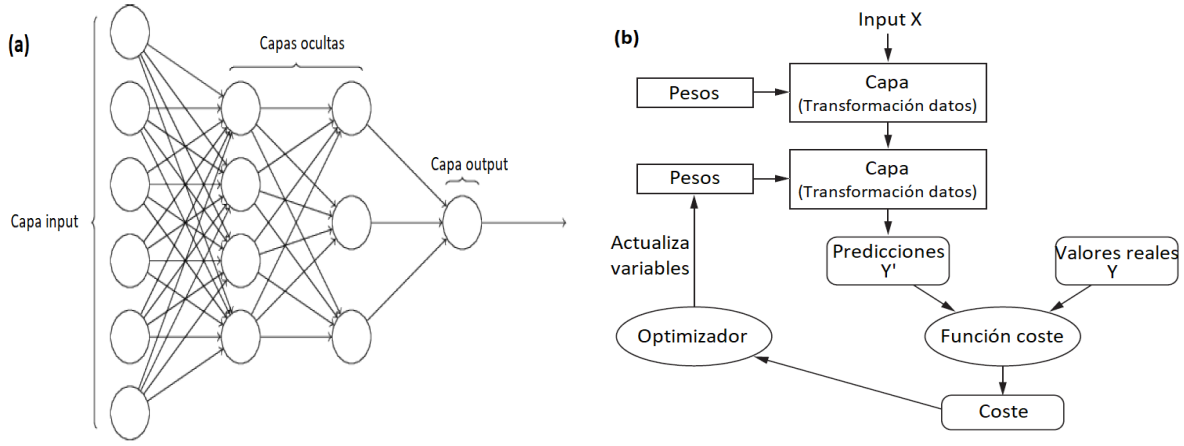

Figura 3: (a) Ejemplo de estructura genérica de una red neuronal estándar. Se pueden observar 3 partes principales, la capa *input*, las ocultas y la *output* [19]. (b) Esquema de los pasos del funcionamiento de una red neuronal estándar [18].

El objetivo de una red neuronal, como se ha mencionado, es realizar una determinada tarea utilizando computación clásica. En ese aspecto no se diferencia de la algorítmica convencional. Lo que realmente supone un cambio de paradigma es la forma en la que una red neuronal hace su tarea y cómo se implementa dicha tarea. Una red neuronal aprende. El proceso de aprendizaje es básicamente de prueba-error. La red toma un *input* (que puede ser un dígito manuscrito) y se espera de ella que responda con un *output* (que sería el valor numérico del dígito). Durante un proceso de entrenamiento a la red se le indica si ha acertado o no, y mediante técnicas que se resumen en los siguientes párrafos, se corrige a la red en el sentido que mejora su aprendizaje. Internamente todos los procesos de aprendizaje se realizan mediante una serie de transformaciones tanto lineales como no lineales, que se llevan a cabo en cada capa de neuronas de su estructura. Las operaciones en el interior de la red modifican los parámetros internos de la red, en cada exposición a los ejemplos.

Como se observa en la figura 3a, la primera capa es la capa de *inputs*, donde se introducen los datos al programa, por lo que tiene que haber el mismo número de neuronas que de datos de entrada. Las siguientes capas son las capas ocultas, estas pueden ser muy variadas, la elección del número de capas ocultas y de neuronas por capa depende del objetivo para el que se emplee la red neuronal. La última capa es la de *outputs*, de nuevo, tiene que haber el mismo número de neuronas que de *outputs* esperados. Las flechas indican que cada neurona envía su respuesta a todas las neuronas de la capa siguiente. En el caso de los *inputs*, se envía cada valor a todas las neuronas de la primera capa oculta y, en estas, a dicho valor se le aplica una transformación a la que llamaremos función de activación. Existen varios tipos de funciones de activación, como la rectificadora ReLU (*Rectified Linear Unit*) y la tangente hiperbólica, pero la más empleada en redes neuronales es la función sigmoidea $\sigma(z)$ [19].

$$\sigma(z) \equiv \frac{1}{1+e^{-z}} \quad con \quad z = -\sum_j w_j x_j - b \tag{3.3}$$

Donde x es el vector de valores que le llega a la neurona de la capa anterior, w es el vector peso

de la neurona (*weights*) y b es la desviación (*bias*). Estos dos últimos parámetros son los parámetros entrenables, característicos de cada neurona, y la función de activación que le asignemos a las neuronas se aplica sobre z. Así, cada neurona elabora su respuesta y la manda a todas las neuronas de la siguiente capa, hasta que se llega a la capa final y se obtiene un resultado. Este se ha de comparar con el resultado que la red neuronal debe aprender a producir. Para ello, se introduce la función de coste, la cual también puede tomar varias formas así que, de nuevo, se introduce la función de error cuadrático medio (ECM), como la más empleada [19].

$$f(w,b) \equiv \frac{1}{2n}\sum_x \|y(x)-a\|^2 \tag{3.4}$$

Siendo n el número total de *inputs*, a el resultado de la red neuronal e $y(x)$ el resultado que cabe esperar y que se busca que la red sea capaz de predecir. Por tanto, el objetivo consiste en buscar unos valores de w y b de modo que $f(x) \approx 0$ y, cuando esto suceda, la red neuronal estará entrenada. En definitiva, el propósito del algoritmo de entrenamiento consiste en minimizar la función coste $f(x)$ en función de los pesos w y la desviación b. Para ello, se emplea el algoritmo de descenso de gradiente.

El objetivo del algoritmo GD consiste en resolver problemas de minimización, y en este caso la función a minimizar es la función coste. El gradiente de una función en un punto de su dominio indica la dirección de máxima variación de la función en dicho punto. De este modo, se calcula el gradiente de la función coste respecto de los parámetros entrenables w y b y, a continuación, se modifica el valor de estos parámetros una pequeña cantidad en la dirección opuesta a la indicada por el gradiente. Esto hace que, al calcular la respuesta con los nuevos valores de w y b, la función de coste se acerque progresivamente a su mínimo. Las modificaciones sobre los parámetros tienen la forma que sigue [19].

$$w_k' = w_k - \eta\frac{\partial f}{\partial w_k} \tag{3.5}$$

$$b_l' = b_l - \eta\frac{\partial f}{\partial b_l} \tag{3.6}$$

Donde η es un parámetro positivo llamado *learning rate*, que indica la magnitud del cambio en los parámetros. El valor que se le asigne a este parámetro depende del problema a resolver pero, generalmente, se trata de que no sea tan grande que no converja al valor del mínimo, ni tan pequeño que cueste llegar al mínimo un tiempo excesivo. Esta es una de las variables llamadas hiperparámetros. Se trata de parámetros que influyen en la red y que, dependiendo del valor que se les asigne, el funcionamiento de esta resulta mejor o peor.

En la práctica, se emplea el algoritmo de descenso de gradiente estocástico SGD (*Stochastic Gradient Descent*) para acelerar el proceso. La idea reside en que para computar el gradiente completo, se han de calcular los gradientes para cada *input* y realizar la media, lo cual, si el número de entradas es muy alto, ralentiza el aprendizaje. Lo que realiza este nuevo algoritmo es calcular el gradiente para una muestra pequeña de *inputs* aleatoria, que recibe el

nombre de *mini-batch*, y promediarlo. Resulta que este promedio proporciona una buena estimación del gradiente total y, a su vez, acelera el proceso de aprendizaje. Por ello, este proceso entra dentro de un bucle que genera los *mini-batches*, el cual, se encuentra dentro de otro bucle, más externo, que corre sobre el número de épocas de entrenamiento. Estas dos variables, el tamaño del *mini-batch* y el número de épocas, también forman parte de los hiperparámetros del sistema.

El algoritmo que se encarga de retroceder por la red cambiando los valores de los parámetros, como se ha indicado, se conoce como algoritmo de de retropropagación (de su nombre en inglés, *Backpropagation Algorithm*), y es el algoritmo central en redes neuronales [18]. Este proceso descrito hasta ahora se puede observar esquemáticamente en la figura 3b. En esta, el optimizador especifica la manera en la que el gradiente de la función coste se emplea para actualizar los valores de los parámetros, es decir, hace referencia al algoritmo SGD. Reproduciendo los pasos del esquema varias veces, para distintos *inputs*, se consigue una red neuronal entrenada.

Por último, es necesario contar con una gran cantidad de datos, es decir, una lista de *inputs* y sus respectivos *outputs*. Esta base de datos se divide en 3 bloques que son, el conjunto de datos de entrenamiento, el de validación y el de examen. Esto es necesario ya que el objetivo no es que la red neuronal se ajuste a los datos de los que se disponen, sino que sepa predecir los resultados. De este modo, la red neuronal lleva a cabo el aprendizaje con el conjunto de datos de entrenamiento, es decir, se realiza todo el proceso para obtener los valores de w y b que minimizan la función de coste solo con este conjunto de datos. A continuación, se emplea el conjunto de validación para ver cómo de bien predice resultados la red para datos que no ha visto antes. Esto sirve para probar distintos valores de los hiperparámetros (η, número de épocas y tamaño del *mini-batch*) y averiguar cuáles ofrecen un mejor funcionamiento de la red. Por último, el conjunto de datos de examen se emplea para comprobar que, tras el entrenamiento y el ajuste de los hiperparámetros, la red funciona adecuadamente y es capaz de predecir respuestas para parámetros de entrada con los que no había trabajado antes, con gran precisión.

3.4. Optimizador clásico GD

Como ya se ha comentado anteriormente, el objetivo del algoritmo *gradient descent* se basa en resolver un problema de minimización, y por eso es útil en este trabajo en el que se busca obtener la máxima absorción posible en el telecom. Para esto, se puede pensar en una función que dependa de la absorción en esta longitud de onda, cuyo mínimo corresponda al máximo de absorción. Una opción para esta función, que es la que se va a emplear, es adaptar la función ECM a este caso, como se indica a continuación.

$$f(\vec{p}) = \frac{1}{2}(f_{NN}(\vec{p}) - 1)^2 \tag{3.7}$$

El vector $\vec{p}$ hace referencia a los parámetros de la estructura de ranuras, es decir, $\vec{p} = (h_{Au}, h_{Mo}, a, p)$. Por tanto, la función, como es lógico, depende de los parámetros que se han de ir cambiando y adaptando para obtener el pico de absorción. En cuanto a la función f_{NN}, esta es la absorción calculada por la red neuronal, de ahí el subíndice NN de *Neural Network*. De este modo, f_{NN}

es un número, la absorción en 1550 nm, que depende de los parámetros de la estructura y, cambiando los valores de estos, este número será más o menos cercano a la unidad. Resulta obvio que la función $f(\vec{p})$ tiene un mínimo para $f_{NN} = 1$, por tanto, al aplicar el algoritmo GD, se puede encontrar ese mínimo.

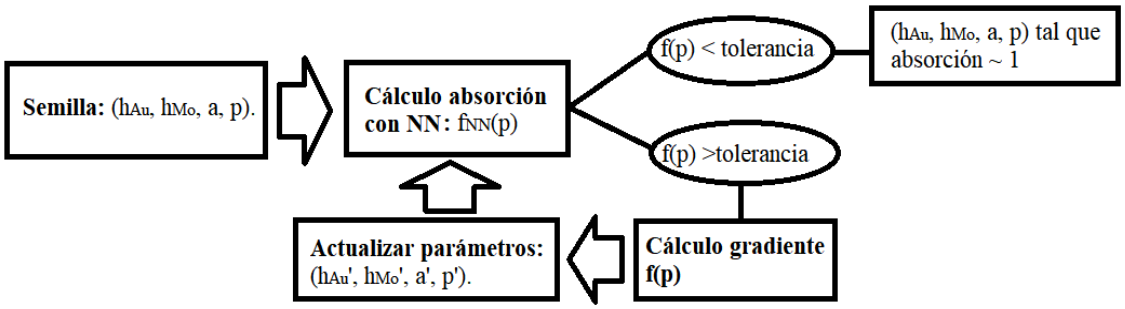

Figura 4: Esquema de los pasos que se siguen en el funcionamiento del algoritmo de optimización *Gradient Descent*, que se va a emplear para encontrar los parámetros geométricos adecuados para la implementación del TES sensible al telecom.

El algoritmo GD lleva a cabo el proceso que se observa en la figura 4. El primer paso en la optimización es elegir un conjunto de parámetros que actúen como "semilla"del optimizador. Cuando no se tiene información del sistema el algoritmo se suele inicializar con una semilla aleatoria. En este estudio, se cuenta con información sobre el comportamiento de las resonancias de absorción y la elección de las semillas se realiza teniendo en cuenta que estas producen un pico de absorción alta cerca de la longitud de onda de interés. Para esta semilla, se calcula, haciendo uso de la red neuronal entrenada, el valor de la absorción en $\lambda = 1550$ nm. Este valor, se sustituye en la función 3.7 y se compara con un valor de tolerancia. Se incluye este valor de tolerancia para no exigir directamente que la absorción sea 1, ya que quizá ese valor resulte inalcanzable. Al inicio, se le adjudica un valor más o menos alto a esta tolerancia, y conforme se comprueba el funcionamiento del programa, se va bajando para obtener la absorción más cercana a 1 posible. Así, si el valor de la función $f(\vec{p})$ es menor que el de la tolerancia, significa que ya se ha llegado a una absorción cercana a la unidad y los parámetros que proporcionan esta absorción son los de la semilla (vía superior en la bifurcación de la figura 4).

Sin embargo, resulta inusual obtener en el primer intento los parámetros que dan máxima absorción por lo que, normalmente, el valor de $f(\vec{p})$ para la semilla es mayor que la tolerancia (vía inferior en la bifurcación de la figura 4). En este caso, se aplica el algoritmo GD y, para empezar, se calcula el gradiente de $f(\vec{p})$ respecto de cada uno de los parámetros h_{Au}, h_{Mo}, a y p. A continuación, se actualizan los valores de estos parámetros del mismo modo que en la red neuronal, mediante las expresiones 3.5 y 3.6, solo que ahora los parámetros de la estructura sustituyen a los parámetros w y b. Como se puede apreciar en dichas expresiones, se cambian los valores de los parámetros en sentido contrario al gradiente de la función y, como el gradiente marca la dirección de máxima variación, así los consecutivos conjuntos de parámetros producen una absorción que minimiza la función $f(\vec{p})$. Con estos nuevos valores de los parámetros más cercanos al mínimo, se vuelve a calcular la absorción mediante la red neuronal y, de nuevo, se

sustituye en la función $f(\vec{p})$ y se compara con la tolerancia. Realizando esto reiteradas veces, cada vez se acercan más los valores de los parámetros a aquellos que proporcionan el mínimo de la función y, por tanto, el máximo en la absorción de 1550 nm.

4. Resultados

4.1. Obtención de los datos de entrenamiento mediante FDTD

Para comenzar a entrenar la red neuronal que se empela en el optimizador GD, se ha de disponer de un conjunto de datos de entrenamiento. Para generarlos, se usa el algoritmo FDTD, al cual se le introducen los parámetros del sistema para que resuelva el espectro correspondiente. De este modo, se tienen los valores de entrada de la red neuronal con sus espectros, los cuales emplea la red para comparar su resultado y aprender. A la hora de generar el set de datos de entrenamiento se limita el rango de parámetros en los que moverse debido a las limitaciones de fabricación del TES comentadas anteriormente. Así, se generan espectros para sistemas con valores de sus parámetros dentro de los siguientes dominios.

- Espesor del oro h_{Au}: [20 nm, 200 nm]
- Espesor del molibdeno h_{Mo}: [20 nm, 200 nm]
- Anchura del agujero a: [0 nm, periodo]
- Periodo p: [1300 nm, 1700 nm]

El código FDTD utilizado no es comercial [17]. El sistema se discretiza con un mallado compuesto por cubos de 5 nm de lado. Teniendo el sistema discretizado a 5 nm, se obtiene un espacio de parámetros extenso. La malla se considera bidimensional, a lo largo de la dirección x se coloca la estructura de ranuras, y a lo largo de la dirección z viaja el haz de luz (ver figura 2). Por tanto, la estructura se coloca, aproximadamente, hacia la mitad de la extensión de la dirección z. Así, en z = 0 se genera el haz de luz, que incide a medio camino con la estructura y, a la salida en punto intermedio entre la estructura y el final de la malla, se analiza cómo se ha modificado el haz al interaccionar con el material.

Se ilumina el sistema con un paquete gaussiano que contiene peso en todas las frecuencias de interés. La iluminación es en incidencia normal y polarización tal que el campo eléctrico oscila perpendicular a la cara de las ranuras (dirección x). Esta elección es importante porque la red de ranuras actúa como un polarizador. Si el campo eléctrico se orienta en la dirección de las ranuras la respuesta del sistema corrugado no difiere de las del sistema sin corrugar.

Se puede observar un ejemplo de resolución del espectro utilizando el algoritmo FDTD en la figura 5a. Este espectro se obtiene para los valores p = 1464 nm, a = 795 nm, h_{Au} = 70 nm y h_{Mo} = 190 nm. En este caso, se aprecia un máximo de absorción en la longitud de onda λ = 1465 nm. Esto corresponde a que, para esa longitud de onda, se da una excitación de un plasmón superficial en la estructura, el cual acaba siendo absorbido por la misma, proporcionando el pico en la absorción que roza la unidad. Puesto que se da esta excitación, se puede observar que la referencia $\lambda \approx p$ sirve para predecir la posición del pico. Por último, en esta imagen también se ve que la curva correspondiente a la transmisión es constante e igual a cero, lo cual se debe a la capa de molibdeno sobre la que se deposita el oro corrugado. Dado que el Molibdeno es opaco en el telecom hace que la transmisión sea despreciable.

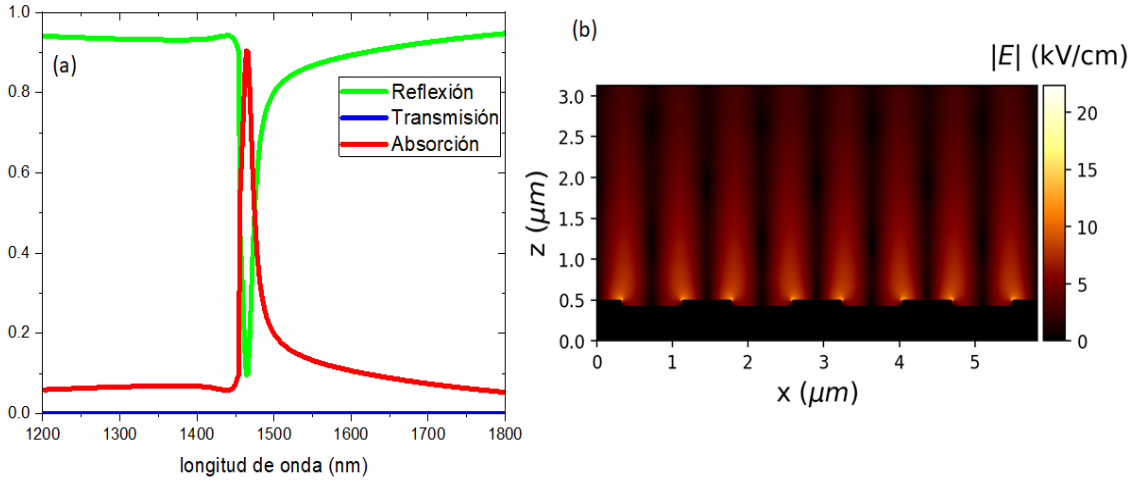

Figura 5: (a) Curvas de absorción, transmisión y reflexión para la nanoestructura de red de ranuras con p = 1455 nm, a = 795 nm, $h_{Au} = 70$ nm y $h_{Mo} = 190$ nm, obtenidas con el algoritmo FDTD en incidencia normal. (b) Módulo del campo eléctrico correspondiente a varias celdas unidad de la nanoestructura, para la longitud de onda del máximo de absorción, $\lambda = 1465$nm.

Por otro lado, la figura 5b muestra el módulo del campo eléctrico correspondiente a varias celdas unidad de la nanoestructura, para la longitud de onda que corresponde al máximo de absorción, $\lambda = 1465$nm. Se observa el patrón de campo característico de un SPP en este tipo de sistemas, en este caso con máximos de campo local en las esquinas de las ranuras. El campo eléctrico es varias veces más intenso que el de la luz incidente, que en este caso es de 1kV/cm, como suele ser habitual en la excitación de plasmones de superficie. Otro dato interesante es que no aparece el típico patrón de reflexión pura que se observa si el metal no está corrugado. La ausencia de franjas de interferencia horizontales es un claro signo de que a esa longitud de onda incidente prácticamente nada de luz se refleja.

Para entrenar la red neuronal, se generan 2600 espectros con el algoritmo FDTD. Es decir, se cuenta finalmente con un conjunto de 2600 valores distintos para los parámetros y sus respectivos espectros, como conjunto de entrenamiento de la red.

4.2. Optimización de la red neuronal

Una vez se cuenta con el conjunto de entrenamiento adecuado, se implementa la red neuronal. Para ello, se emplea *Keras*, que consiste en un entorno de trabajo de *deep-learning* de *Python*, que ofrece una vía para definir y entrenar casi cualquier tipo de modelo de *deep-learning*. Hay que tener en cuenta que esta librería no maneja operaciones como la manipulación de tensores. En cambio, se basa en librearías de tensores especializadas y optimizadas para ello. En este caso, se emplea *TensorFlow* como librería especializada para el cáclulo tensorial [18].

Con este entorno de trabajo, en primer lugar, únicamente es necesario detenerse en definir correctamente el conjunto de datos de entrenamiento, adaptando los datos de entrada y

de salida a tensores con los que pueda trabajar el modelo de red neuronal *TensorFlow-Keras*. A continuación, se ha de definir la estructura de la red neuronal y una de las ventajas que ofrece este entorno es que se pueden definir varias redes neuronales en un mismo programa, e ir eligiendo una u otra según convenga. A cada red neuronal se le llama modelo y se definen dentro del mismo todos los parámetros del sistema, el optimizador, la función de coste, la inicialización de los parámetros... Además, permite definir cada capa de neuronas, independientemente, indicando el número de neuronas de entrada, de salida y la función de activación de las mismas. De este modo, además de elegir los valores de los hiperparámetros, también se elige uno de los modelos implementados. Por último, se entrena la red mediante la función fit() de la librería e iterando sobre el conjunto de datos de entrenamiento.

Durante este proceso se prueban distintos diseños de red neuronal. En cada modelo se cambia el número de capas ocultas, el de neuronas por capa y sus funciones de activación o de coste, entre otras cosas, pero todos tienen en común el número de neuronas de entrada y de salida. En este caso, la capa *input* tiene 9 neuronas, correspondientes a los 4 parámetros geométricos (h_{Au}, h_{Mo}, a y p), a las constantes dieléctricas de los materiales, que se dividen en parte real e imaginaria (ϵ^i_{Au}, ϵ^r_{Au}, ϵ^i_{Mo}, ϵ^r_{Mo}) y a la longitud de onda λ. Se incluye la longitud de onda ya que la red neuronal va a aprender a predecir la reflexión y absorción para una longitud de onda determinada. La capa *output* consiste en 2 neuronas que proporcionan la absorción y reflexión para una determinada longitud de onda. De este modo, dados los 9 parámetros de *input* que definen un estado electromagnético de la nanoestructura para una geometría dada, la red neuronal ha de aprender a devolver como *output* la reflexión y absorción correspondientes.

Para realizar distintos modelos, ya se ha comentado que dos de los aspectos que se pueden variar son la función de coste y la de activación. En cuanto a la función de coste, a pesar de que la más sencilla sea la de coste cuadrático, esta se puede cambiar por la función de coste de entropía cruzada (*cross-entropy cost function).* En cuanto a la de activación, aunque hay varias, lo más común es probar, además de la sigmoide, la ReLU y la tangente hiperbólica. Estas funciones tienen las siguientes formas.

$$Cross-entropy: \ f(w,b) = -\frac{1}{n}\sum_x [ylna + (1-y)ln(1-a)] \tag{4.1}$$

$$ReLU: \ f(z) = max(0,z) \tag{4.2}$$

$$Tanh: \ f(z) = \frac{2}{1+e^{-2x}} - 1 \tag{4.3}$$

La *cross-entropy function* tiene la virtud de que, cuando la diferencia entre la respuesta de la red y la real es muy grande, lo cual es muy fácil que suceda al inicio, más rápido aprende la red, y esto es lo que se busca. En cuanto a las funciones de activación, simplemente hay que probar cuál es la que mejor se adapta al problema, no hay una forma de elegir adecuada salvo probando.

Otro aspecto a variar en estos modelos es la estructura de la red. Hay una infinidad de combinaciones de número de capas ocultas y de neuronas por capa a elegir pero, en general, la

tendencia es que a mayor número de capas ocultas, menor se hace el coste, y a mayor número de neuronas por capa, más rápido aprende la red [20].

Por último, para definir un modelo se pueden escoger dos caminos, usar *Sequential class* o *functional API* [18]. El modelo secuencial, es decir, el primero, sirve únicamente para elaborar pilas lineales de capas de neuronas, mientras que, el segundo, se emplea para elaborar gráficos acíclicos dirigidos de las capas, permitiendo construir estructuras arbitrarias. En este caso, se emplea el primero, ya que no se van a realizar topologías complicadas de la red, solamente capas lineales de neuronas unidas entre ellas.

De este modo, se implementan varios modelos, combinando el uso de distintas funciones de activación y de coste y de distintas estructuras de la red. Tras varias pruebas, finalmente se encuentra que aquel que ofrece mejores resultados es el que se describe a continuación. La estructura de la red consiste en 2 capas ocultas de 40 neuronas cada una, más las capas de entrada y de salida, todas densas, lo que en este entorno significa que están completamente conectadas. A todas ellas, y a la capa *output*, se les aplica como función de activación la sigmoide y, como función de coste se emplea la *binary cross-entropy*. Además, para evaluar el buen funcionamiento de la red, se emplea la función *mean squared error*. Por último, como optimizador emplea el *stochastic gradient descent*, ya mencionado anteriormente.

Una vez elegido el modelo de red neuronal más eficiente, se optimiza su funcionamiento probando distintos valores de los hiperparámetros del sistema. Ya se ha mencionado anteriormente que una buena elección de los hiperparámetros del sistema es esencial para un buen funcionamiento de la red neuronal y, que para ir variando los valores de estos parámetros se evalúa su efecto en el conjunto de datos de validación. Por ello, tras cada cambio que se haga en la red, se evalúan la función de coste y la función ECM, empleada como métrica, tanto en los datos de entrenamiento como en los de validación. El objetivo consiste en que ambas evolucionen de la manera más rápida posible hacia el cero. Sin embargo, no hay ninguna regla que indique cómo elegir los valores de los hiperparámetros, simplemente hay que probar y a menudo esta es una de las tareas más costosas y en la que se ha centrado mucho el estudio de las redes neuronales.

Finalmente, a través de este proceso de optimización, se llega a los siguientes valores óptimos: 50 épocas, *mini-batch size* = 64, *learning rate* $\eta = 1.0$, parámetro de regularización $\lambda = 0.1$, momento = 0.45, *decay* $= 1 \cdot 10^{-6}$ y *nesterov* = True. El momento, junto con el *decay* y el *nesterov*, son parámetros característicos del optimizador SGD que ofrece este entorno de trabajo *Keras*, para mejorar el proceso de convergencia hacia el mínimo de la función a optimizar [21]. Cabe resaltar que estos valores asignados a los hiperparámetros son óptimos para el funcionamiento de esta red en concreto. Para cada red que se implemente con una función distinta, se tiene que realizar el proceso de optimización para hallar los valores adecuados de los parámetros en ese caso concreto.

Por último, esta red se entrena con los 2400 espectros de los 2600 generados mediante el FDTD y se dejan 100 para validación, que son los empleados para el ajuste de los hiperparámetros, y otros 100 como conjunto de espectros de examen. Por tanto, en total se tienen 2·2600 espectros, ya que se calcula transmisión y reflexión, y cada uno de ellos, contiene información

de 300 longitudes de onda.

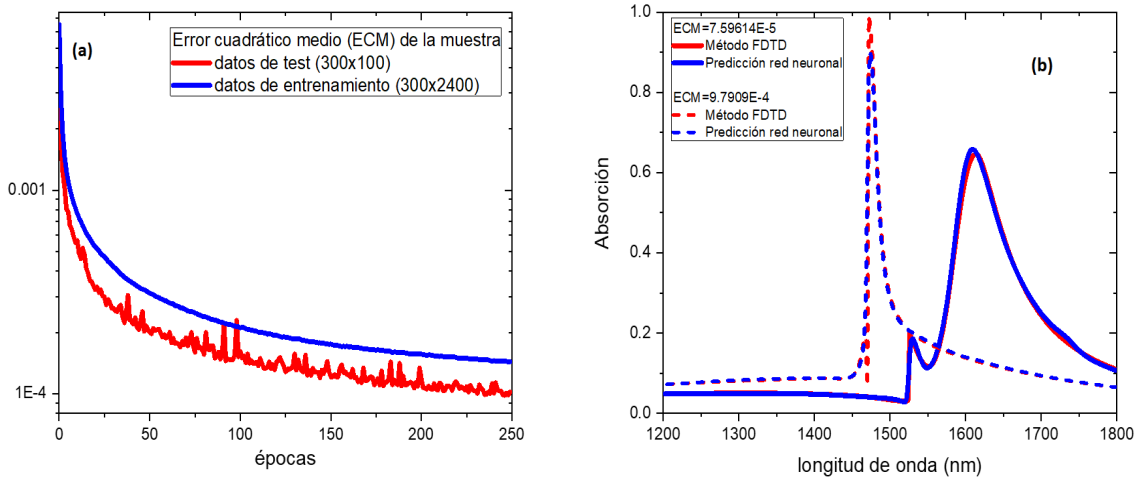

Figura 6: (a) Función ECM frente al número de épocas para los datos de entrenamiento (curva azul) y los de test (curva roja). Esta se obtiene mediante la red neuronal, implementada en el entorno *Keras*, descrita en el texto. (b) Espectros obtenidos, por parte de la red neuronal, para 2 casos concretos del conjunto de datos del test. El de mayor ECM corresponde a p = 1525 nm, a = 615 nm, h_{Au} = 145 nm, h_{Mo} = 150 nm, y el de menor a p = 1470 nm, a = 940 nm, h_{Au} = 115 nm, h_{Mo} = 100 nm. Se observa la comparación entre las predicciones obtenidas mediante la red neuronal (curvas azules) y los resultados del algoritmo FDTD (curvas rojas), apreciando un mejor ajuste entre estas para el caso con menor ECM.

En la figura 6a se observa la evolución de la función ECM frente al número de épocas para el conjunto de los datos de entrenamiento y para los de test. En ambos, se aprecia como esta función llega a valores muy cercanos al cero, consiguiendo incluso valores aún más bajos para el conjunto de datos de test. Con estos valores de esta función, que se encarga de medir la correcta actuación de la red neuronal, se consigue reproducir con bastante precisión los espectros obtenidos mediante el algoritmo FDTD.

En la figura 6b, se muestra la resolución del espectro para estos dos casos concretos, que se obtienen con la red neuronal, con los datos del conjunto para test, y se comparan con los calculados con FDTD. Esta comparación evidencia el hecho de que a menor ECM, con mayor precisión resuelve los espectros la red neuronal, ya que para el caso con ECM = $7.60{\cdot}10^{-5}$, se aprecia una mayor coincidencia con el resultado numérico del algoritmo FDTD, que para el caso con ECM = $9.80{\cdot}10^{-4}$. En este segundo caso, el resultado de la red neuronal no termina de ajustarse al correspondiente al algoritmo FDTD, de modo que predice la posición del pico existente, pero con una menor precisión, ya que no llega a dar el valor exacto de absorción que se alcanza. Por tanto, la red neuronal tiene cierto margen de error, pero el grado de concordancia en la mayoría de las resoluciones de espectros que resuelve la red es muy alto, por lo que se puede concluir que esta funciona correctamente.

Por tanto, se tiene una red que aprende en poco tiempo y resuelve la absorción de manera muy precisa, es decir, se ha alcanzado el objetivo de funcionamiento de la red. A continuación, se emplea este red neuronal entrenada como parte del algoritmo de optimización, con el fin de hallar los parámetros adecuados para que la nanoestructura presente un máximo de absorción en la longitud de onda de 1550 nm. Sin embargo, aunque la red constituya un método muy eficiente para el proceso de optimización, una vez obtenido un diseño óptimo, resulta necesario validarlo con los cálculos del algoritmo FDTD.

4.3. Optimización de los parámetros geométricos del TES

Con la red neuronal entrenada ya se tiene la herramienta principal del funcionamiento del algoritmo de optimización *Gradient Descent*. El programa de optimización se prueba para varias semillas, que en este caso son parámetros geométricos, y distintos valores de tolerancia y del parámetro η de las ecuaciones 3.5 y 3.6, que mide cómo de grande es el cambio que se realiza en los parámetros en la fase de actualización.

Es importante resaltar el papel esencial de la red neuronal en este proceso en cuanto a la velocidad de cálculo. La red empleada para este proceso ha requerido calcular 2600 espectros. Este conjunto de datos de entrenamiento se ha obtenido de manera numérica mediante el algoritmo FDTD y para ello, se han tenido 20 ordenadores trabajando durante 30 días. Por tanto, para generar los 2600 espectros con un único ordenador son necesarios 600 días. Sin embargo, la red neuronal es un factor 10^6 más rápida que el FDTD realizando los mismos cálculos. Con este algoritmo de optimización, para cada semilla es necesario un número distinto de iteraciones, pero se han empleado del orden de 1000 iteraciones hasta alcanzar un valor de la absorción cercano a la unidad. De este modo, teniendo en cuenta que en cada iteración se tiene que calcular la absorción para evaluar $f(\vec{p})$ y compararla con el valor de tolerancia, si se emplea el algoritmo FDTD tardaría 6 meses en realizar 1000 iteraciones. Además, durante este proceso de optimización resulta preciso hacer pruebas con varias semillas, para comprobar la posible existencia de distintas soluciones al problema. Para ello, se implementa un bucle en el programa que, una vez terminadas las iteraciones con una semilla, perturba la semilla inicial creando otra nueva y reiniciando el proceso. Cada cálculo de los que se han realizado se hace con unas 50 semillas, lo cual da lugar a 50000 iteraciones, que con el FDTD equivaldría aproximadamente a 31 años de cálculo. La mejora es patente, con la red neuronal entrenada se tarda un cuarto de hora aproximadamente en realizar el mismo número de cálculos.

Una vez implementado el optimizador con el algoritmo GD, se realizan las pruebas. Como se ha comentado, se toman semillas aleatorias diferentes, con el objetivo de ver qué posibles estructuras distintas pueden dar esta condición de resonancia. En este proceso se ha de tener en cuenta que no todas las estructuras que se obtengan pueden ser luego implementadas ya que, dependiendo de la temperatura de operación y de la capacidad calorífica que ofrezcan, pueden, o no, ser útiles. Por tanto, para cada estructura optimizada, se realiza una estimación de la temperatura crítica de operación y la capacidad calorífica asociadas, siendo que la primera depende de los espesores de Au y Mo, y la segunda de la propia temperatura crítica y el tamaño del TES. Para ello, el tamaño lateral del dispositivo se toma como el número de ranuras por el

periodo, que en este caso supone 20·p.

Con esto, se puede obtener un rango de valores de los parámetros para los que se tiene el pico de absorción en $\lambda = 1550$ nm, y elegir el más adecuado de acuerdo a los requerimientos físicos y de fabricación del TES.

De este modo, los resultados son los siguientes. Se obtienen 2 rangos de parámetros como solución de la minimización realizada por parte del algoritmo GD. Esto se observa en la figura 7 donde se puede apreciar que, tanto los valores de los parámetros optimizados como los de la temperatura crítica y la capacidad calorífica asociadas, se agrupan en rangos entorno a absorciones algo menores de 0.91 y 0.95.

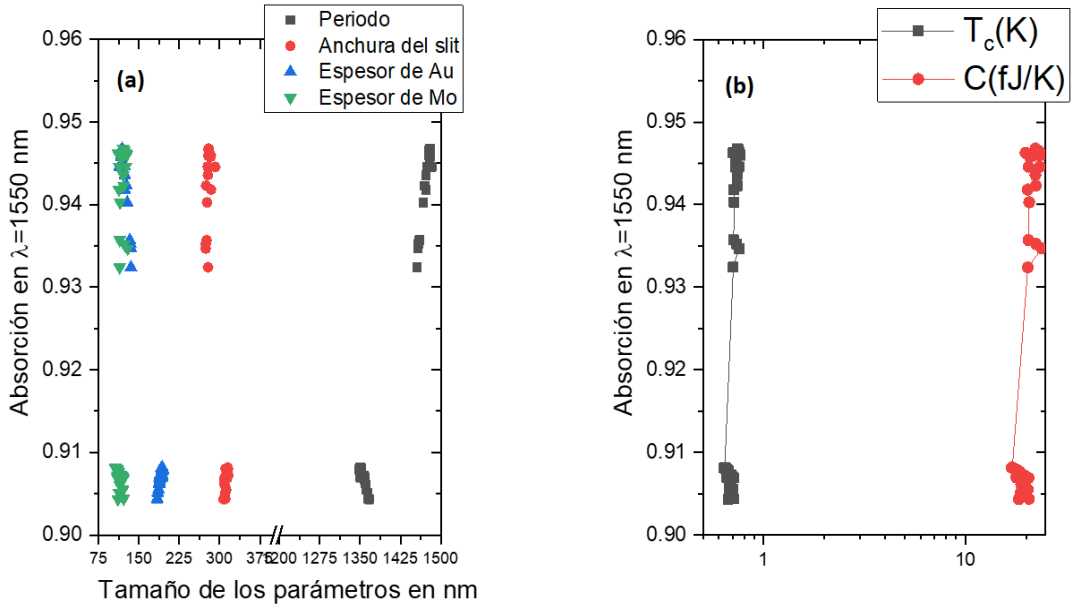

Figura 7: Resultados de absorción para la optimización de los parámetros geométricos del TES sensible al telecom, mediante el algoritmo *Gradient Descent* que utiliza la red neuronal entrenada. Para su obtención, se emplean 50 semillas distintas y en cada una de ellas se realizan 1000 iteraciones. (a) Absorción frente a los valores de los parámetros optimizados para el diseño del TES sensible a $\lambda = 1550$ nm. (b) Absorción resultante frente a los valores de temperatura crítica de operación y capacidad calorífica, que se estima que tiene el TES sensible a $\lambda = 1550$ nm que se busca.

Del conjunto de parámetros que ofrecen estos dos rangos observados en la figura 7, se hace una selección de aquellos que ofrecen un buen balance entre absorción alta, temperatura crítica y capacidad calorífica razonables para el funcionamiento del TES. Estos resultan buenos candidatos a implementar el TES sensible a la radiación de longitud de onda de 1550 nm. Sin embargo, antes de entrar en detalle con estos resultados, se ha de recordar que los espectros correspondientes a estos conjuntos de parámetros se han obtenido mediante el cálculo de una red neuronal. Por tanto, se tiene que comprobar que los espectros que arrojan estos parámetros

sean de la forma que se busca (pico de absorción en 1550 nm) introduciendo en el algoritmo FDTD el valor de los mismos.

De los resultados que muestra la figura 7, el mejor resultado en términos de absorción se obtiene para un sistema de ranuras con p = 1480 nm, a = 280 nm, $h_{Au} = 120$ nm y $h_{Mo} = 115$ nm. Estos valores se obtienen de redondear los que devuelve el optimizador y que mejor encajan con el discretizado utilizado en el método FDTD (5 nm). Se ha comprobado que el cambio en la curva de absorción es mínimo al pasar de valores nominales obtenidos por el optimizador a los valores usados en los cálculos FDTD. Como se puede ver en la figura 8, el espectro que predice la red neuronal encaja muy bien con el obtenido con el método FDTD. La absorción que se alcanza con esta configuración es de aproximadamente un 98 %. Es esta estructura la que, finalmente, se propone en este trabajo como un sistema para construir un TES sensible a la radiación de longitud de onda de 1550 nm.

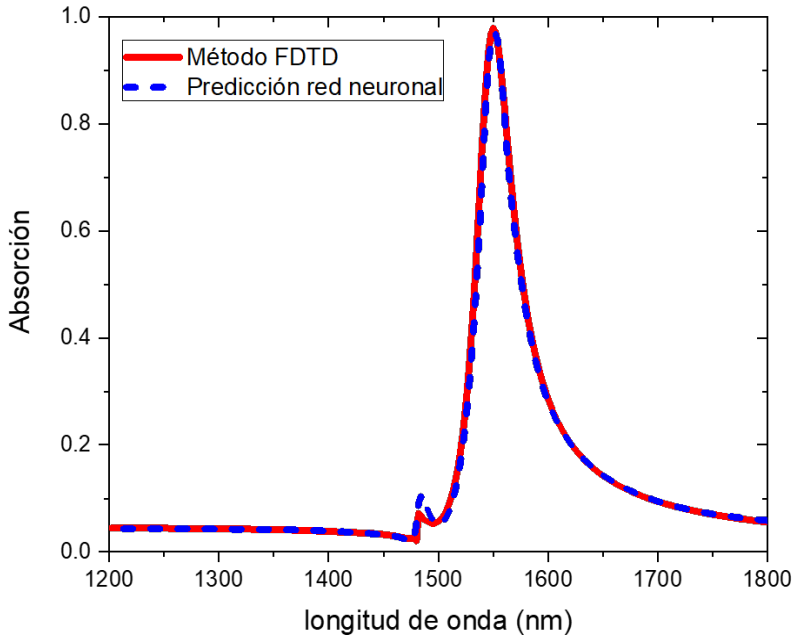

Figura 8: Espectros de absorción calculados con el método FDTD (línea roja) y la red neuronal (línea azul) para los parámetros p = 1480 nm, a = 280 nm, $h_{Au} = 120$ nm y $h_{Mo} = 115$ nm. Estos parámetros han sido obtenidos durante el proceso de optimización y constituyen el resultado principal del trabajo.

Para verificar que estos parámetros corresponden a un diseño de TES que pueda ofrecer un funcionamiento adecuado, se analizan los valores obtenidos para la temperatura crítica de operación y la capacidad calorífica. Como ya se ha expuesto anteriormente, para tener una ΔT medible, son necesarios valores de la capacidad calorífica cercanos a 10fJ/K. En este caso, para estos valores de los parámetros, se obtiene que C = 21.01 fJ/K, por lo que se encuentra en el orden de magnitud esperado. En cuanto a la temperatura T_c, la fórmula empleada para obtenerla no es más que una estimación. Se trata de una fórmula extraída de ajustar datos experimentales

de los TES de los que ya se dispone, cuyos espesores de oro y molibdeno son distintos a los que se han obtenido. La única forma de asegurar el valor de la temperatura de operación es experimentalmente, pero con esta estimación se puede formar una idea acerca de la funcionalidad del TES implementado con estos parámetros. Así, se obtiene $T_c = 720$ mK, una temperatura mayor que los 100 mK a los que operan los TES diseñados para rayos X. Sin embargo, dado que la C obtenida es razonable, se puede construir un diseño válido. Por lo tanto, se concluye que la combinación de los valores de T_c y C es adecuada para un buen funcionamiento del TES.

Por último, se comprueba que el valor que se obtiene de la absorción no sea muy sensible a la variación de los parámetros del TES. Esto es necesario ya que puede suceder que, a la hora de fabricar el TES, no sea posible construirlo con los valores exactos de los parámetros obtenidos con el optimizador, debido a las limitaciones de la fabricación o por resolución. Con el algoritmo FDTD se ha constatado que un cambio de $\pm$ 5 nm/10 nm en el espesor del oro o en la anchura del periodo, provoca un corrimiento de la posición del pico en, aproximadamente, la misma cantidad. Esto, dependiendo de la aplicación puede tener consecuencias más o menos importantes, por ejemplo, en el caso de que los fotones a medir no sean estrictamente monocromáticos, este problema puede no resultar tan grave.

Por todo esto, los valores de los parámetros: p = 1480 nm, a = 280 nm, $h_{Au} = 120$ nm y $h_{Mo} = 115$ nm, que proporcionan una absorción del 98 % de la luz incidente en incidencia normal para $\lambda = 1550$ nm, establecen un diseño prometedor para implementar el TES sensible a la radiación del rango de las telecomunicaciones, y se completa, de esta manera, el objetivo del trabajo. La validación definitiva del funcionamiento de un TES con este diseño requiere medidas sobre dispositivos reales, lo cual puede llevar a modificar el diseño. No obstante, las herramientas desarrolladas a lo largo de este trabajo van a permitir emplear la información obtenida experimentalmente para encontrar nuevos diseños de dispositivos óptimos. Por otro lado, este trabajo se ha centrado en obtener diseños en una longitud de onda fija. Sin embargo, los TES permiten medir la energía del fotón incidente y para algunas aplicaciones, como la astronomía, se podría contemplar el sacrificar algo de eficiencia de absorción para conseguir espectros más anchos que permitan medir estos directamente. De este modo, de nuevo, las herramientas desarrolladas en este trabajo, pueden adaptarse para optimizar estos nuevos diseños.

5. Conclusiones

El objetivo de este trabaja consistía en diseñar un TES con una absorción optimizada en el rango telecom ($\lambda = 1550$ nm), excitando SPPs mediante las nanoestructuras adecuadas. Para lograr esto ha sido necesario el desarrollo y la optimización de una plataforma basada en redes neuronales, que acelerase el cálculo. Como resultado, se ha conseguido implementar y entrenar una red neuronal que permite una aceleración del cálculo en un factor 10^6. Finalmente, mediante el uso de esta herramienta, se ha propuesto un diseño preliminar que alcanza una eficiencia en la absorción del 98 % a la frecuencia de interés.

El correcto funcionamiento de este primer diseño se ha de verificar experimentalmente. No obstante, la red neuronal desarrollada permitirá seguir mejorando este resultado ya que, por ejemplo, la frecuencia de la resonancia en este caso puede ser muy sensible a pequeñas variaciones de los parámetros geométricos. Por tanto, un siguiente paso a este trabajo puede ser tratar de reducir la sensibilidad a las variaciones en los parámetros de fabricación. En este sentido, este primer diseño está además basado en una geometría fija, pero puede resultar interesante explorar modificaciones en la propia geometría. Esto constituye otro aspecto que se podría estudiar, por ejemplo, variando la profundidad de las ranuras en el oro, o con una forma de nanoestructura diferente a las ranuras.

Por tanto, se puede concluir que este trabajo ha sido una buena primera aproximación al problema, donde se ha obtenido la propuesta de diseño funcional del TES y una red neuronal entrenada para seguir explorando mejoras en esta. Las opciones para seguir afinando esta propuesta son varias, pero en cualquier caso, el paso final consistirá en elaborarlo y comprobar su funcionamiento experimentalmente.

Por último, también se pueden investigar otras posibles aplicaciones. Para la aplicación propuesta en este trabajo, se ha optimizado la respuesta del TES a una longitud de onda fija. Sin embargo, los TES son sensores capaces de medir la energía del fotón incidente, por lo que modificando de manera adecuada el diseño, podrían proponerse estructuras sensibles a un rango mayor de longitudes de onda. De este modo, se abre la posibilidad de realizar espectroscopía en aplicaciones como, por ejemplo, la astronomía. De nuevo, la red neuronal elaborada en este estudio podría entrenarse para tratar de aplicarla a este nuevo rango de aplicaciones. En definitiva, los resultados aquí alcanzados pueden constituir el punto de partida de futuros trabajos en este campo.

Referencias

[1] B. Cabrera, R. M. Clarke, P. Colling, A. J. Miller, S. Nam, and R. W. Romani. Detection of single infrared, optical, and ultraviolet photons using superconducting transition edge sensors. *Applied Physics Letters*, 73(6):735–737, aug 1998.

[2] D.Rosenberg, A.E.Lita, A.J.Miller, S.Nam, and R.E.Schwall. Performance of photon-number resolving transition-edge sensors with integrated 1550 nm resonant cavities. *IEEE Transactions on Appiled Superconductivity*, 15(2):575–578, jun 2005.

[3] Carlos Pobes, Lourdes Fabrega, Agustin Camon, Nieves Casan-Pastor, Pavel Strichovanec, Javier Sese, Javier Moral-Vico, and Rosa Maria Jaudenes Calleja. Development of cryogenic x-ray detectors based on mo/au transition edge sensors. *IEEE Transactions on Applied Superconductivity*, 27(4):1–5, jun 2017.

[4] The x-ray integral field unit (x-IFU) for athena. In Tadayuki Takahashi, Jan-Willem A. den Herder, and Mark Bautz, editors, *Space Telescopes and Instrumentation 2014: Ultraviolet to Gamma Ray*. SPIE, jul 2014.

[5] Thomas Gerrits, Adriana Lita, Brice Calkins, and Sae Woo Nam. Superconducting transition edge sensors for quantum optics. *Superconducting Devices in Quantum Optics, Quantum Science and Technology*, 2016.

[6] William L. Barnes, Alain Dereux, and Thomas W. Ebbesen. Surface plasmon subwavelength optics. *Nature*, 424(6950):824–830, aug 2003.

[7] R. H. Ritchie. Plasma losses by fast electrons in thin films. *Physical Review*, 106(5):874–881, jun 1957.

[8] Allen Taflove and Susan C Hagness. *Computational electrodynamics: the finite-difference time-domain method; 3rd ed.* Artech House antennas and propagation library. Artech House, Boston, MA, 2005.

[9] J.Peurifoy, Y.Shen, L.Jing, Y.Yang, F.Cano-Renteria, B.G.DeLacy, J.D.Joannopoulos, M.Tegmark, and M.Soljacic. Nanophotonic particle simulation and inverse design using artificial neural networks. *SCIENCE ADVANCES*, 2018.

[10] John M.Martinis, G.C.Hilton, K.D.Irwin, and D.A.Wollman. Calculation of TC in a normal-superconductor bilayer using the microscopic-based usadel theory. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 444(1-2):23–27, apr 2000.

[11] Quantum materials and devices. url: https://www.qmad.es/.

[12] F. Hao and P. Nordlander. Efficient dielectric function for FDTD simulation of the optical properties of silver and gold nanoparticles. *Chemical Physics Letters*, 446(1-3):115–118, sep 2007.

[13] Wolfgang S. M. Werner, Kathrin Glantschnig, and Claudia Ambrosch-Draxl. Optical constants and inelastic electron-scattering data for 17 elemental metals. *Journal of Physical and Chemical Reference Data*, 38(4):1013–1092, dec 2009.

[14] H. A. Bethe. Theory of diffraction by small holes. *Physical Review*, 66(7-8):163–182, oct 1944.

[15] T. W. Ebbesen, H. J. Lezec, H. F. Ghaemi, T. Thio, and P. A. Wolff. Extraordinary optical transmission through sub-wavelength hole arrays. *Nature vol*, 391(6668):667–669, feb 1998.

[16] Sergio G. Rodrigo, Fernando de Leon-Perez, and Luis Martin-Moreno. Extraordinary optical transmission: fundamentals and applications. *Proceedings of the IEEE*, 104(12):2288–2306, dec 2016.

[17] Sergio G. Rodrigo. *Optical properties of nanostructured metallic systems*. Springer Berlin Heidelberg, 2012.

[18] Francois Chollet. *Deep learning with python*. Manning, 2017.

[19] Michael Nielsen. *Neural networks and deep learning*. Springer, 2013.

[20] Dianjing Liu, Yixuan Tan, Erfan Khoram, and Zongfu Yu. Training deep neural networks for the inverse design of nanophotonic structures. *ACS Photonics*, 5(4):1365–1369, feb 2018.

[21] TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.