ISSN 1870-4069

Sesgos inductivos relacionales en mecanismos de atención

Víctor Mijangos [1], Ximena Gutierrez-Vasques [2], Verónica E. Arriola [1], Ulises Rodríguez-Domínguez [1], Alexis Cervantes [1], José Luis Almanzara [1]

[1] Universidad Nacional Autonoma de Mexico,
Facultad de Ciencias,
México

[2] Universidad Nacional Autonoma de Mexico,
CEIICH,
México

{vmijangosc, v.arriola, ulises.rodriguez.dominguez, alexis.cervantes, jose-luis}@ciencias.unam.mx, xim@unam.mx

Resumen. El aprendizaje inductivo busca construir modelos generales a partir de ejemplos específicos, siendo guiado por sesgos inductivos que influyen en la selección de hipótesis y determinan la capacidad de generalización. En este trabajo, nos centramos en caracterizar los sesgos inductivos relacionales presentes en los mecanismos de atención, entendidos como suposiciones sobre las relaciones subyacentes entre los datos. Desde el marco del aprendizaje profundo geométrico, analizamos los mecanismos de atención más comunes en términos de sus propiedades de equivariancia respecto a subgrupos de permutaciones, lo que nos permite proponer una clasificación basada en sus sesgos relacionales.

Palabras clave: Mecanismos de atención, transformadores, sesgos inductivos, aprendizaje profundo geométrico.

Relational Inductive Biases in Neural Attention Mechanisms

Abstract. Inductive learning aims to build general models from specific examples, guided by inductive biases that influence hypothesis selection and determine generalization capability. In this work, we focus on characterizing relational inductive biases present in attention mechanisms, understood as assumptions about underlying relationships between data elements. Within the framework of geometric deep learning, we analyze common attention mechanisms in terms of their equivariance properties with respect to permutation subgroups, which allows us to propose a classification based on their relational biases.

Keywords: Attention mechanisms, transformers, inductive biases, geometric deep learning.

1. Introducción

Uno de los paradigmas más comunes en los métodos actuales de aprendizaje de máquina es el *aprendizaje inductivo* cuyo objetivo es construir una función general, o hipótesis, a partir de ejemplos particulares observados [20,21]. De esta manera, dado un conjunto de ejemplos de entrenamiento (pares de entrada-salida), el modelo busca una hipótesis, dentro de un espacio de posibles hipótesis, que se ajuste bien a los datos de entrenamiento y que pueda generalizarse bien a instancias que no se han visto antes.

Las suposiciones realizadas por un algoritmo de aprendizaje para proponer hipótesis, que estén definidas para todo el dominio del problema y no sólo para los valores de los ejemplares observados, constituyen un *sesgo inductivo*. Son estas suposiciones las que le dotan de potencial para generalizar a datos no vistos [21]. Otra forma de sesgo inductivo son aquellas suposiciones que prefieren ciertas hipótesis sobre otras. Por lo tanto, los sesgos inductivos juegan un papel clave en la capacidad de generalización de los modelos de aprendizaje de máquina.

A pesar de que los sesgos inductivos son uno de los componentes que permiten el aprendizaje de diversas tareas, e.g., procesamiento del lenguaje natural (NLP), reconocimiento y generación de imágenes, es difícil encontrar trabajos que aborden formalmente los fundamentos de los sesgos inductivos, particularmente en el aprendizaje profundo.

En este artículo nos centraremos en exponer el funcionamiento de uno de los tipos de sesgos inductivos más prominentes, los *sesgos relacionales*, que son aquellos que explotan la estructura relacional inherente a los datos [3]. Nuestro análisis estará aplicado a las redes neuronales de tipo transformador (*transformer*) y sus mecanismos de atención. Proponemos adoptar las estrategias del aprendizaje profundo geométrico para describir los sesgos relacionales, pues proveen un marco de estudio robusto para modelar estructuras relacionales en los datos, así como transformaciones y simetrías.

2. Marco teórico

2.1. Sesgos inductivos relacionales

Los sesgos inductivos son suposiciones previas, *a priori*, que ayudan a que el aprendizaje pueda elegir mejor una hipótesis sobre otra [20]. Tom Mitchell [21] propone plantearse la pregunta de cuáles son los *a priori* necesarios para que el algoritmo de aprendizaje pueda realizar un proceso deductivo para generalizar sobre una nueva instancia x. Estas suposiciones *a priori* son precisamente lo que se entiende por sesgos inductivos.

Definición 1 (Sesgo inductivo) *Sea $f^* : X \to Y$ una función objetivo arbitraria y sea $D = \{(x, f^*(x)) : x \in X\}$ un conjunto de datos de entrenamiento y A un algoritmo de aprendizaje. Un sesgo inductivo de A es un conjunto mínimo de suposiciones B tal que para todo $x \in X$:*[1]

$$(B \wedge D \wedge x) \vdash A(x, D)$$

[1] Usamos la notación de Mitchell [21] tomando a B y D como conjunciones sobre sus elementos vistos como proposiciones.

. Donde $A(x,D)$ es la predicción del algoritmo entrenado sobre D en la instancia x, $\vdash$ denota la inferencia de $A(x,D)$ a partir de $(B \wedge D \wedge x)$:

En particular, nos interesan los sesgos inductivos relacionales [3] para analizar datos conformados por conjuntos de entidades, que se encuentran estructuralmente relacionadas entre sí. Por ejemplo, sea X un conjunto de enunciados en español, cada ejemplar x está constituido por el conjunto de palabras $\{x_1,...,x_n\}$ en el enunciado.

Definición 2 (Sesgo inductivo relacional) *Dado un conjunto de entidades que conforman un ejemplar $x = \{x_1,...,x_n\}$, un sesgo inductivo relacional es una suposición acerca de las relaciones entre dichas entidades. Es decir, es una estructura relacional (x,G), tal que $G = (V,E)$ es una gráfica relacionando los elementos de x.*

2.2. Aprendizaje profundo geométrico

Para estudiar los sesgos inductivos relacionales dentro de los **mecanismos de atención**, es crucial adoptar un marco teórico que permita caracterizar y relacionar los diferentes mecanismos que se han propuesto. Bronstein et al. [6] adoptan un enfoque basado en características geométricas y sus grupos de simetrías. Papillon et al. [22] han extendido este enfoque a una perspectiva topológica. Gavranović et al [13] proponen un marco más general basado en teoría de categorías para englobar tanto perspectivas geométricas y topológicas (que llaman "descendentes" o *top-down*) como marcos "ascendentes" (*bottom-up*), que parten de la construcción de arquitecturas con base a métodos de diferenciación automática [1], [23], [5]. Nos enfocaremos en las relaciones establecidas por una gráfica que se asume dentro de los modelos de atención. Adoptamos la teoría del aprendizaje geométrico profundo [7], ya que este modelo teórico, al enfocarse en estructuras geométricas del dominio de datos, es ideal para expresar las relaciones entre las entidades de los datos.

El aprendizaje profundo geométrico adopta una metodología de estudio basada en el programa de Felix Klein para la geometría [16]: se define un conjunto o dominio de elementos y grupos de transformaciones asociadas a este dominio. Con base en esta idea, el objetivo del aprendizaje geométrico profundo es determinar características (principalmente relacionales) sobre el dominio de datos y determinar el tipo de capas ocultas que pueden aprovechar dichas características.

La idea esencial detrás del aprendizaje profundo geométrico radica en estudiar a los ejemplares x y las relaciones entre sus entidades x_i. Estas suelen representarse a través de gráficas $G = (V,E)$ donde los vértices en V se asocian a las entidades y E representa sus relaciones. El dominio de los datos y la tarea determinan el tipo de capas de un modelo profundo. Estas capas deben preservar la información relevante para resolver las tareas. Para formalizar esto, es primordial el concepto de función equivariante.[2]

[2] Un ejemplo de especial interés son las capas convolucionales. En estas, el dominio consiste en imágenes, que pueden entenderse como elementos $x \in \mathbb{R}^{H\times W\times C}$, donde H es la altura, W la anchura y C el número de canales. La acción de una traslación $g \in \mathscr{G}$ puede representarse por una matriz de traslación $T = \rho(g)$, de tal forma que Tx es una traslación de la imagen x. Una convolución $f : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{H\times W\times C}$ es equivariante ante traslaciones en tanto se puede comprobar la igualdad $f(Tx) = Tf(x)$.

Definición 3 (Función equivariante) *Sea $\mathscr{G}$ un grupo de simetrías sobre un conjunto X, una función $f : X \to X$ se dice que es $\mathscr{G}$-equivariante si para toda acción del grupo $g \in \mathscr{G}$ se cumple que:*

$$f\big(\rho(g)x\big) = \rho(g)f(x). \tag{1}$$

Donde $\rho(g)$ es la representación (matricial) de la acción g.

Otro concepto importante dentro del marco teórico es el de invarianza y las funciones invariantes donde se tiene que $f\big(\rho(g)x\big) = f(x)$. Dadas las características de los mecanismos de atención nos enfocaremos en funciones $\mathscr{G}$-equivariantes, o simplemente equivariantes. El marco del aprendizaje profundo geométrico permite definir una gran cantidad de capas principalmente para arquitecturas de redes neuronales de gráficas. Veremos que las capas de atención caben bajo el concepto de una capa gráfica.

Definición 4 (Capa gráfica) *Una capa gráfica (o de envío de mensajes) depende de una gráfica $G = (V,E)$ con un sistema de vecindades $\mathscr{N}_v$ para cada vértice asociado a una entidad x_v. De tal forma que la nueva entidad oculta h_v es de la forma:*

$$h_v = \phi\left(x_v, \bigoplus_{u \in \mathscr{N}_v} \psi(x_v, x_u)\right). \tag{2}$$

En donde se distinguen los siguientes elementos:

1. *Función de mensaje $\psi(x_v, x_u)$: Genera un mensaje (generalmente un vector) con base en x_v y sus vecinos x_u.*
2. *Función de agregación $\bigoplus$: Determina cómo se combinan los mensajes para la actualización; se pide que sea un operador conmutativo.*
3. *Función de actualización $\phi(\cdot,\cdot)$: Define la forma final que tendrá la nueva representación h_v con base en la representación original x_v y la estructura relacional.*

En la Sección 3, estudiamos las equivarianzas de los mecanismos de atención a partir de subgrupos de permutaciones finitas S_n. Para esto tomamos como punto de partida el teorema de Cayley que apunta a que todo grupo es isomorfo a un subgrupo de permutaciones [8].

2.3. Mecanismos de atención

Los mecanismos de atención fueron introducidos, para el problema de traducción automática en redes recurrentes, por Bahdanau et al. [2]. Posteriormente, Vaswani et al. [27] propusieron reemplazar las recurrencias por mecanismos de atención que trabajaran sobre los mismos datos de entrada, lo que llamó auto-atención (*self-attention*). A partir de estos trabajos se han definido nuevas arquitecturas y capas de atención, de las cuales aquí presentamos las más comunes, comenzando por la auto-atención [27].

Definición 5 (Auto-atención) *Dado un conjunto de entidades de un ejemplar* $\{x_1,..,x_n\}$ *con* $x_i \in \mathbb{R}^d$*, un mecanismo de auto-atención es una capa de la forma:*

$$h_i = \sum_j \alpha(x_i,x_j)\psi_v(x_j), \tag{3}$$

donde $\alpha(x_i,x_j)$ *son los pesos de atención definidos como:*

$$\alpha(x_i,x_j) = Softmax\Big(\frac{\psi_k(x_j)^T\psi_q(x_i)}{\sqrt{d}}\Big). \tag{4}$$

Denotamos con ψ_q, ψ_k *y* ψ_v *a las proyecciones aplicadas a las entidades de entrada, generalmente definidas por una función lineal o afín.*

Los transformadores [27] integran un tipo de atención similar a la de Bahdanau [2] que representa las entidades de entrada (en el codificador) con las de la salida (el decodificador). Esta atención codificador-decodificador, puede ser definida bajo los conceptos de la Definición 5 restringiendo la forma en que se determinan las relaciones.

Definición 6 (Atención codificador-decodificador) *Supóngase una bipartición determinada por los conjuntos de entidades* $X = \{x_1,...,x_m\}$ *e* $Y = \{y_1,...,y_n\}$*. Definimos la atención codificador-decodificador como el mecanismo que para todo* y_i*, con* $i \in \{1,...,n\}$*, obtiene las representaciones ocultas como:*

$$h_i = \sum_{j=1}^{m} \alpha(y_i,x_j)\psi_v(x_j). \tag{5}$$

Donde $\alpha(y_i,x_j)$ *se define de forma similar a la Ecuación 4.*

Cuando los mecanismos de atención se utilizan para predecir un elemento (entidad), dado un conjunto previo (como en la generación de texto), el entrenamiento de estos modelos no puede asumir que los elementos previos dependen de los futuros, que aún no conoce. Para lidiar con esto, los decodificadores en los transformadores implementan mecanismos de atención enmascarada [27].

Definición 7 (Atención enmascarada) *Dado un conjunto de entrada de entidades ordenadas* $x_1,..,x_n$*, un mecanismo de atención enmascarada es una capa que obtiene representaciones de la forma:*

$$h_i = \sum_{j\leq i} \alpha(x_i,x_j)\psi_v(x_j), \tag{6}$$

donde los pesos de atención $\alpha(x_i,x_j)$ *se estiman como en la Ecuación 4.*

En la definición previa hemos definido un orden para simplificar la expresión en la sumatoria. La idea de desconectar nodos dentro de la gráfica que define las relaciones entre las entidades en un mecanismo de atención puede llevarse todavía más lejos. Por ejemplo, Child et al. [9] sugieren un mecanismo de atención por pasos donde las relaciones están acotadas por distintas restricciones, como un límite a los elementos previos con los que se puede conectar una entidad.

Definición 8 (Atención por pasos) *La atención por pasos es un tipo de atención dispersa donde, dada una entrada con entidades ordenadas $x_1, ..., x_n$, se obtienen sus representaciones como:*

$$h_i = \sum_{t \leq j \leq i} \alpha(x_i, x_j) \psi_v(x_j), \tag{7}$$

donde $t = \max\{0, i-k\}$ para una constante k y $\alpha(x_i, x_j)$ es como en la Ecuación 4.

Otra forma general de obtener mecanismos de atención es asumir que las relaciones entre las entidades de cada ejemplar son arbitrarias, y están definidas por la matriz de adyacencia de una gráfica [28].

Definición 9 (Atención en gráficas) *Dadas las entidades $\{x_1, .., x_n\}$ de un dato con información sobre sus vecinos $\mathcal{N}_i$ para toda i, un mecanismo de atención en gráficas es una capa de la forma:*

$$h_i = \sum_{j \in \mathcal{N}_i} \alpha(x_i, x_j) \psi_v(x_j), \tag{8}$$

donde $\alpha(x_i, x_j)$ son los pesos de atención como en la Ecuación 4.

La atención dispersa y la atención en gráficas se diferencian en que la atención dispersa asume relaciones arbitrarias para todos los ejemplares del dominio de datos, mientras que en la atención en gráficas, las relaciones dependen de cada dato particular. Este último mecanismo es la forma más general de un mecanismo de atención de donde se pueden derivar los mecanismos definidos anteriormente.

2.4. Atención como núcleo generalizado

Tsai et al. [26] proponen clasificar los mecanismos de atención con base en un núcleo (*kernel*), centrado en la función $\alpha(x_i, x_j)$ de la Ecuación 4. Bajo esta perspectiva, los pesos de atención dependen de una función $k : X \times X \to \mathbb{R}$, donde X es el espacio de rasgos para los mecanismos de atención (tanto rasgos posicionales como no posicionales [27]). En este marco, $k(\cdot, \cdot)$ es un núcleo, que en los mecanismos de atención es exponencial, y ya que las proyecciones ψ_q, ψ_k no son simétricas en general, se considera a k como un núcleo generalizado (no simétrico).

La visión basada en kenerls [26] no considera los sesgos inductivos que pueden existir dentro de estos mecanismos. En lo que sigue, asumimos que los pesos de atención $\alpha(x_i, x_j)$ están determinados por el softmax de los productos $k(x_i, x_j)$ bajo un núcleo generalizado. Esto es:

$$\alpha(x_i, x_j) = Softmax_K\big(K_{i,j}\big), \tag{9}$$

No abordamos a profundidad las consecuencias que tiene el uso de diferentes tipos de núcleos [26], puesto que nuestra propuesta se enfoca a sesgos inductivos relacionales. Aunque cabe señalar que esta visión daría pie a sesgos relacionales no necesariamente binarios. Esta es una perspectiva que dejamos como trabajo a futuro. Señalamos que nuestra propuesta no es contraria, sino complementaria a la visión de estos mecanismos en el marco de los núcleos generalizados.

3. Resultados

A partir de la revisión de los diferentes mecanismos de atención (Sección 2.3) se pudo observar que estos mecanismos comparten un núcleo general en donde se tiene una agregación de forma aditiva de los valores de entidades vecinas ponderados por una probabilidad generalmente estimada por una función softmax. Este hecho ya introduce un sesgo inductivo relacional dentro de los mecanismos de atención.

Proposición 1 (Sesgo de relaciones estocásticas) *Las capas de atención asumen que las relaciones entre las entidades $x_1,...,x_n$ son estocásticas.*

La matriz de atención, que denotaremos como α, define la matriz estocástica asociada, en donde cada entrada $\alpha_{i,j} := \alpha(x_i,x_j) = p(x_j|x_i)$. Esta probabilidad está dada en términos de la función softmax (Ecuación 9).

Cabe señalar que los mecanismos de atención se enmarcan dentro de capas de la forma expuesta en la Ecuación 2 (Definición 4): i) la función de mensaje está determinada como $\psi(x_i,x_j) = \alpha(x_i,x_j)\psi_v(x_j)$; ii) la agregación se da por medio de la suma sobre todas las entidades de un dato; y iii) la función de actualización obtiene las nuevas representaciones como la agregación de los mensajes. Resaltando el marco gráfico y estocástico, proponemos una definición general de capa de atención.

Definición 10 (Capa de atención) *Una capa de atención es un tipo de capa gráfica (Definición 4) que estima la representación de un conjunto de entidades $\{x_1,x_2,...,x_n\}$ con sistema de vecindades $\{\mathcal{N}_i : i = 1,2,...,n\}$ a partir del valor esperado sobre una distribución p de estas vecindades:*

$$h_i = \mathbb{E}_{p\sim\mathcal{N}_i}\big[\psi_v(x)\big],$$

Donde la esperanza se estima sobre las relaciones de una matriz de adyacencia dada como (Ecuación 9):

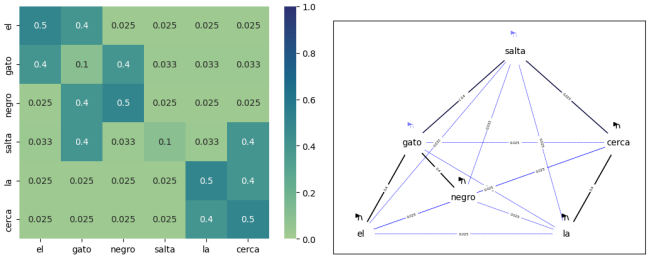
$$\alpha(x_i,x_j) = Softmax_K(K_{i,j}).$$

$\psi_v(x)$ es una proyección de los datos sobre el espacio de valores.

En lo que sigue, nos basamos en esta definición para mostrar los sesgos inductivos relacionales a los que responde cada uno de los mecanismos de atención. En particular, nos enfocaremos en la forma en que las relaciones se manifiestan en la matriz de atención, pues, como hemos señalado, es esta matriz la que determina las relaciones entre entidades. Para esto, nos centramos en los procesos de enmascaramiento.

Lemma 1. *Las relaciones que subyacen al dominio de datos en un mecanismo de atención, se manifiestan en la matriz de atención como un proceso de enmascaramiento.*

Demostración. El proceso de enmascaramiento consiste en eliminar ciertas entradas de la matriz de atención α. Sea K la matriz de productos (Ecuación 9) al que se aplica la función softmax. Una entrada se enmascara a partir de la asignación $K_{i,j} = -\infty$ antes de aplicar el softmax, de tal forma que se tiene que $\alpha_{i,j} = 0$. Claramente esto representa



(a) Matriz de atención (b) Gráfica relacional

Fig. 1. Ejemplo de las relaciones establecidas por medio de un mecanismo de auto-atención.

una desconexión en la estructura gráfica relacional. Si $G = (V,E)$, podemos fácilmente definir el proceso de enmascaramiento como:

$$K_{i,j} = \begin{cases} k(x_i,x_j) & \text{si } (i,j) \in E \\ -\infty & \text{si } (i,j) \notin E. \end{cases}$$

De esta forma, el enmascaramiento está determinado por el tipo de relaciones que se asumen en el dominio de datos.

Teorema 1 (Sesgo relacional en auto-atención) *Las capas de auto-atención asumen un sesgo inductivo relacional con base en una gráfica completamente conectada [6].*

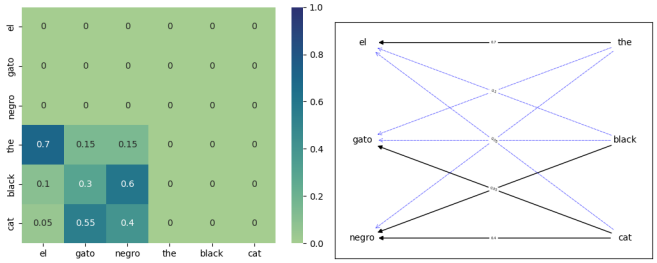
Demostración. En una capa de auto-atención (Definición 5), la agregación, que en las redes de transmisión de mensajes se aplica sobre los vecinos en la gráfica del nodo que representa a la entidad x_i, se aplica sobre todas las entidades de entrada.

Con base en el Lema 1, si K es la matriz de productos, es claro que para todo x_i, x_j entidades de los datos de entrada se tiene que $K_{i,j} = k(x_i,x_j) > -\infty$, por lo que en la matriz de atención $\alpha_{i,j} > 0$. Esto implica que no se realizan desconexiones en la gráfica subyacente; es decir, se asume una gráfica completamente conectada.

Por tanto, las capas de auto-atención asume un sesgo inductivo relacional donde las entidades de un datos se relacionan todas entre sí, esto es, $E = X \times X$ (véase la Figura 1). Con respecto a las capas de atención codificador-decodificador (Definición 6), éstas se conforman a partir de una agregación con respecto a las entidades de entrada y buscan únicamente representar entidades de salida.

Teorema 2 (Sesgo relacional atención codificador-decodificador) *Las capas de atención codificador-decodificador asumen un sesgo inductivo relacional con base en una gráfica bipartita.*

Demostración. En una capa de atención codificador decodificador se tiene un conjunto de datos $x_1, x_2, ..., x_n, y_1, y_2, ..., y_m$ con una partición de las entidades $X = \{x_1, ..., x_n\}$ y



(a) Matriz de atención (b) Gráfica relacional

Fig. 2. Ejemplo de las relaciones establecidas por medio de un mecanismo de atención codificador decodificador.

$Y = \{y_1, ..., y_m\}$ donde $X \cap Y = \emptyset$. La matriz de productos K entonces está determinada de la siguiente forma:

$$K_{i,j} = \begin{cases} k(y_i, x_j) \text{ si } y_i \in Y \wedge x_j \in X \\ -\infty \text{ en otro caso.} \end{cases}$$

Esto es, $\alpha_{i,j} > 0$ si y sólo si $x_i \in X$ y $y_j \in Y$, por lo que la matriz de atención sólo conecta el bloque inferior izquierdo (Figura 2).

La partición está determinada por las entidades del dato de entrada y del dato de salida; generalmente, se trata de una gráfica dirigida que va de las entradas hacia las salidas. Esta atención se puede aplicar tanto a transformadores como a redes recurrentes.

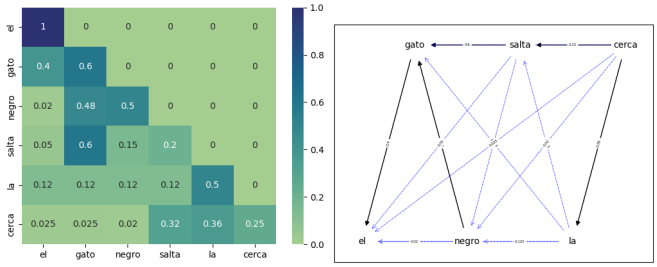
El tercer mecanismo a revisar es la atención enmascarada (Definición 7). Estos tipos de atención asumen un orden en las entidades de entrada, de tal forma que, al pensarse como una gráfica, una conexión se da entre dos nodos si y sólo si una de las entidades precede a otra en este orden.

Teorema 3 (Sesgo relacional en atención enmascarada) *Las capas de atención enmascarada asumen un sesgo inductivo relacional de orden total.*

Demostración. Sea $x = \{x_1, x_2, ..., x_n\}$ el conjunto de entidades de entrada, y defínase un orden $O(x) = \{(x_i, x_{i+1}) : i = 1, ..., n-1\}$ sobre las entidades. Al tomar la matriz de productos K del mecanismo de atención, consideramos el enmascaramiento determinado por las entradas de la matriz en la siguiente forma:

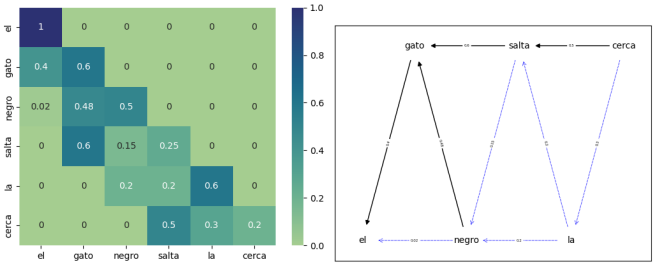
$$K_{i,j} = \begin{cases} k(x_i, x_j) & \text{si } j \leq i \\ -\infty & \text{si } j > i. \end{cases}$$

De tal forma que la matriz de atención tendrá entradas $\alpha_{i,j} \neq 0$ si y sólo si el elemento x_j precede al elemento x_i en $O(x)$, mientras que las otras entradas representan desconexiones en la gráfica. Claramente, las relaciones diferentes de 0 están en la parte triangular superior de la matriz de atención (Figura 3).



(a) Matriz de atención (b) Gráfica relacional

Fig. 3. Ejemplo de las relaciones establecidas por medio de un mecanismo de atención enmascarada.



(a) Matriz de atención (b) Gráfica relacional

Fig. 4. Ejemplo de las relaciones establecidas por medio de un mecanismo de atención por pasos.

Ya que se cumple: i) para toda i, $i \leq i$ por identidad; ii) si $j \leq i$ y $i \leq j$ entonces $i = j$ (de lo contrario habría conexiones en la parte triangular superior de la matriz de adyacencia); y iii) cuando $k \leq j$ y $j \leq i$, entonces $k \leq i$ pues i se conecta con todos los precedentes, podemos concluir que el orden es total.

El tipo de relaciones en las capas de atención enmascarada introducen un sesgo inductivo relacional donde se asume que las entidades no se conectan con elementos subsecuentes. Esto, como es bien sabido, es útil para la representación de secuencias.

Tanto la atención enmascarada como la atención dispersa por pasos (Definición 8) pueden verse como formas de relacionar las entidades de una gráfica únicamente con elementos previos. Esto define una gráfica dirigida cuya matriz de adyacencia cuenta únicamente con la parte triangular inferior. En el caso de la atención por pasos, también se presentan ceros en la pare inferior de la matriz de adyacencia pero acotado a un número dado de elementos previos (véase Figura 4).

Teorema 4 (Sesgo relacional en atención por pasos) *Las capas de atención por pasos asumen un sesgo inductivo donde un elemento se conecta con los p elementos previos para un p fijo y dado un orden previo.*

Demostración. Se puede observar que la atención por pasos es similar a la atención enmascarada y también requieren de un orden $O(x)$ sobre las entidades. Bajo este orden, se pueden definir las entradas de K como:

$$K_{i,j} = \begin{cases} k(x_i, x_j) & \text{si } t \leq j \leq i \\ -\infty & \text{si } t > j > i \end{cases}$$

Definimos aquí $t = \max\{0, i-p\}$ para algún $p \geq 0$. Las relaciones entre las entidades sólo se dan con los t elementos previos, lo que define una gráfica acíclica dirigida (DAG por sus siglas en inglés) que determina el sesgo inductivo relacional en este tipo de mecanismos.

3.1. Sensibilidad a equivarianzas

Para analizar el tipo de equivarianzas ante las cuales son sensibles las capas de atención, nos basamos en subgrupos de permutaciones [8]. Con respecto a las capas de auto-atención tenemos el siguiente resultado ya presentado en [6].

Teorema 5 (Equivarianza en auto-atención) *Las capas de auto-atención son equivariantes ante permutaciones.*

En la atención enmascarada y la atención por pasos, podemos ver que existe una dependencia de un orden previamente establecido.

Teorema 6 (Equivarianza en atención enmascarada) *Las capas de atención enmascarada son equivariantes ante traslaciones.*

Demostración. Una traslación es una función $\sigma(i) = i + m$ para alguna m constante. Como hemos mostrado, las capas de atención enmascarada asumen un orden $O(x)$ sobre las entidades de entrada para crear el enmascarameinto de la matriz de atención α con base en la matriz de productos K. Bajo la traslación es fácil observar que la definición de cada entrada de K:

$$K_{\sigma(i),\sigma(j)} = \begin{cases} k(x_{\sigma(i)}, x_{\sigma(j)}) & \text{si } \sigma(j) \leq \sigma(i) \\ -\infty & \text{si } \sigma(j) > \sigma(i) \end{cases}$$

preserva el orden (si $j \leq i$ entonces $\sigma(i) \leq \sigma(j)$).

Teorema 7 (Equivarianza en atención por pasos) *Las capas de atención por pasos son equivariantes ante traslaciones.*

Demostración. El Teorema 4 introdujo un orden sobre las entidades que, sin embargo, no define relaciones de orden total, pues tenemos que para un p fijo si $j < i - p$ entonces la entidad j no se relaciona con la entidad i, aunque sí puede relacionarse con entidades relacionadas con i, por lo que no tenemos una relación transitiva. Sin embargo, como hemos señalado, la atención por pasos define una gráfica acíclica dirigida (DAG) en donde podemos definir un único orden topológico que hace de esta gráfica un orden total. Por tanto, al igual que con la atención enmascarada, una traslación no modifica las relaciones, pues el orden total resultante con la traslación se preserva.

Corolario 1 *Los sesgos inductivos relacionales de las capas de atención por pasos engloban los de la atención enmascarada.*

La atención por pasos generaliza a la atención enmascarada en tanto basta escoger $t = \text{máx}\{n : n = |x|\}$, es decir, considerar los elementos previos como el valor máximo que puede tener una secuencia.

Teorema 8 (Equivarianza en atención codificador-decodificador) *La atención codificador decodificador es equivariante ante permutaciones en bloques.*

Demostración. Una permutación en bloque es un epimorfismo $\sigma : X \to X$ tal que, dada una relación de equivalencia $\sim$, se tiene que si $x \sim y$ en X, entonces $\sigma(x) \sim \sigma(y)$. Como se señaló en el Teorema 2, la atención codificador-decodificador asume una partición $X \cup Y$ de los datos, de tal forma, que existe una relación de equivalencia $\sim$ que subyace a dicha partición. Claramente si $x_i, x_j \in X$ entonces $x_i \sim x_j$ por lo que $\sigma(x_i) \sim \sigma(x_j)$ y, por tanto, $\sigma(x_i)$ y $\sigma(x_j)$ siguen siendo elementos de X. De forma similar, las permutaciones sobre elementos de Y permanecen en Y. Además los conjuntos X e Y siguen siendo disjuntos ante este tipo de permutaciones, por lo que la partición prevalece.

3.2. Discusión

Proponemos una jerarquía de mecanismos de atención comunes, basada en los sesgos relacionales. Otros trabajos han caracterizado a los transformadores de acuerdo al tipo de arquitectura en que se presentan [18], [19] sin entrar en detalles en las características de la atención. El trabajo de Tsai et al. [26], por su parte, presenta una metodología para la elección de mecanismos de atención, aunque no aborda la forma en que los diferentes tipos de enmascaramiento se relacionan o qué tipo de sesgos introducen.

El Cuadro 1 muestra la clasificación de los mecanismos de atención desde el más general al más particular con base al sesgo relacional y a las equivarianzas que presentan. La atención en gráficas es más general que la atención dispersa. La auto-atención relaciona todas las entidades entre sí, mientras que la atención enmascarada sólo permite relaciones con elementos previos, y la atención por pasos pone una cota al número de elementos previos. Finalmente, la atención codificador-decodificador requiere una bipartición.[3]

La tabla previa también incluye el tipo de datos que estos modelos podrían trabajar. La atención en gráficas es útil para conjuntos de datos donde cada dato tiene una estructura gráfica independiente; por ejemplo, se ha usado para aplicaciones de clasificación de nodos, así como para superficies 3d descritas por triangulaciones; un ejemplo de esto el es modelo de GAT [28]. Las redes de auto-atención suelen usarse en modelos del lenguaje auto-codificados como BERT [11], donde se asume una relación bidireccional. Por su parte, la atención enmascarada se utiliza en modelos auto-regresivos como GPT [10] en los que se asume que se desconoce los elementos siguientes, por lo que sólo hay relaciones hacia atrás. La atención dispersa es utilizada

[3] La implementación de diferentes capas de atención y de transformadores puede encontrarse en: https://victormijangosdelacruz.github.io/MecanismosAtencion/.

Tabla 1. Clasificación de los mecanismos de atención más comunes.

Mecanismo de atención	Relaciones	Grupo de simetrías	Tipo de datos
Atención gráfica	Depende del dato	Dinámica	Con estructuras gráficas por cada dato
Atención dispersa	Arbitraria	Arbitraria	Con relación gráfica específica
Auto-atención	Completamente conectada	Permutaciones	Relacionales o bidireccionales
Enmascarada	elementos previos	Traslación	Secuenciales
Por pasos	p elementos previos	Traslación	Secuenciales acotados
Codificador-decodificador	Bipartición	Sistema de bloques	Particionales

por modelos como Unlimiformer [4] el cuál procesa bloques de texto (acotando en base a la cercanía) y se aplica en clasificación y análisis textuales. Modelos como ViT [12][17] o S4 [14] utilizan atención por pasos en el procesamiento de imágenes como secuencias de tókens, transformando secciones de la imagen o parches. Finalmente T5 [24] incorpora atención codificador-decodificador entre la partición de los datos en elementos de entrada y de salida. Se ha aplicado al lenguaje para traducción o resumen automático. El Cuadro 2 resume estas arquitecturas y sus aplicaciones.

Una de las principales limitantes de los transformadores es que el funcionamiento de estos radica en mecanismos de atención que justamente asumen gráficas completamente conectadas, requiriendo de una gran cantidad de datos para poder aprender relaciones adecuadas. Agregar restricciones a las relaciones que se pueden presentar introduce un sesgo relacional que puede ayudar a mejorar el rendimiento de las arquitecturas basadas en estas capas, siempre y cuando estos sesgos inductivos respondan a la estructura de los datos.

4. Conclusiones y trabajo a futuro

En este trabajo hemos presentado una aproximación teórica a los sesgos inductivos relacionales dentro de los mecanismos de atención. Para esto, nos hemos basado en la teoría del aprendizaje profundo geométrico [6] que nos ha permitido estudiar el tipo de subgrupos de permutaciones ante los cuales éstos mecanismos son equivariantes. Hemos así conformado una clasificación según la suposición que cada mecanismo hace sobre las relaciones subyacentes en las entidades del dominio de los datos. Nuestro análisis proporciona una comprensión más profunda de cómo funcionan los mecanismos de atención y cómo pueden procesar fenómenos complejos, como variaciones sintácticas en lenguas naturales.

Asimismo, extendimos la caracterización de Tsai et al. [26] del enmascaramiento y las relaciones en los mecanismos de atención. Sin embargo, en los mecanismos de atención las relaciones se dan no sólo entre pares de entidades (una gráfica común), sino que pueden existir relaciones de mayor orden (tripletas, cuadripletas, etc.). Las no linealidades en estos mecanismos, como las evidentes desde la perspectiva de los kernels generalizados [26], pueden ser una fuente para dichas relaciones de mayor

Tabla 2. Modelos con uso de atención para lenguaje y visión.

Modelo	**Tipo de Atención**	**Aplicaciones/Efectos**
GAT [28]	En gráficas	Clasificación de nodos, clustering, sistemas de recomendación, modelos 3d.
BERT [11]	Auto-atención	Análisis de sentimiento, clasificación de texto, preguntas y respuestas.
GPT [**?**]	Enmascarada	Generación de texto, traducción automática, resumen automático. Permite el pre-entrenamiento generativo de un modelo grande de lenguaje a través de múltiples tareas de manera no supervisada.
Unlimiformer [4]	Atención dispersa	Análisis de texto, clasificación de texto, procesamiento de lenguaje natural. Permite aprender secuencias de largo alcance con recursos computacionales limitados.
ViT [12]	Por pasos	Clasificación de imágenes, detección de objetos, segmentación de imágenes. Captura patrones de atención espaciales tanto agnósticos (e.g., en primeras capas) como sensibles (e.g., en capas más profundas) al contenido.
T5 [24]	Codificador decodificador	Traducción automática, resumen automático, generación de texto. Permite el entrenamiento del modelo a través de múltiples tareas de manera no supervisada.

orden. Como ejemplo, en el trabajo reciente de [25] se pudo establecer que los transformadores visuales consideran interacciones espaciales de alto orden dentro de cada bloque básico de sus capas. En [15] interpretan la interacción entre tókens de texto en un modelo de transformador, construyendo relaciones jerárquicas y mostrando que éstas interacciones no se dan a pares sino que toman lugar a órdenes mayores.

Como trabajo futuro resulta un reto investigar a fondo las relaciones de mayor orden que pueden darse en éstos mecanismos para poder tener una caracterización completa de los mismos. Otro aspecto relevante es investigar las relaciones que existen con otro tipo de capas como las convolucionales, que pueden verse como un subconjunto de las capas de atención [6]. Ésto puede proporcionar una comprensión más completa de los mecanismos de atención y su papel en las arquitecturas de transformadores.

Agradecimientos. Agradecemos los comentarios de los revisores que ayudarán a la calidad del presente trabajo. También queremos agradecer a los proyectos PAPIIT TA100924 y TA100725 de la UNAM.

Referencias

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467, (2016)
2. Bahdanau, D.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, (2014)

3. Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al.: Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261, (2018)
4. Bertsch, A., Alon, U., Neubig, G., Gormley, M. R.: Unlimiformer: Long-range transformers with unlimited length input (2023), https://arxiv.org/abs/2305.01625
5. Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., et al.: Jax: composable transformations of python+ numpy programs, (2018)
6. Bronstein, M. M., Bruna, J., Cohen, T., Veličković, P.: Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. arXiv preprint arXiv:2104.13478, (2021)
7. Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., Vandergheynst, P.: Geometric deep learning: going beyond euclidean data. IEEE Signal Processing Magazine, vol. 34, no. 4, pp. 18–42 (2017)
8. Cayley, A.: On the theory of groups, as depending on the symbolic equation θn= 1. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 7, no. 42, pp. 40–47 (1854)
9. Child, R., Gray, S., Radford, A., Sutskever, I.: Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509, (2019)
10. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., Salakhutdinov, R.: Transformer-XL: Attentive language models beyond a fixed-length context. In: Korhonen, A., Traum, D., Màrquez, L. (eds) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 2978–2988. Association for Computational Linguistics (2019)
11. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021), https://arxiv.org/abs/2010.11929
13. Gavranović, B., Lessard, P., Dudzik, A., von Glehn, T., Araújo, J. G. M., Veličković, P.: Position: Categorical deep learning is an algebraic theory of all architectures (2024), https://arxiv.org/abs/2402.15332
14. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces (2022), https://arxiv.org/abs/2111.00396
15. Hao, Y., Dong, L., Wei, F., Xu, K.: Self-attention attribution: Interpreting information interactions inside transformer. In: The Thirty-Fifth AAAI Conference on Artificial Intelligence. AAAI Press (2021)
16. Kisil, V. V.: Erlangen program at large: an overview. Advances in applied analysis, pp. 1–94 (2012)
17. Li, Y., Wang, J., Dai, X., Wang, L., Yeh, C.-C. M., Zheng, Y., Zhang, W., Ma, K.-L.: How does attention work in vision transformers? a visual analytics attempt. IEEE Transactions on Visualization and Computer Graphics, vol. 29, no. 6, pp. 2888–2900 (2023)
18. Liu, Y., Zhang, Y., Wang, Y., Hou, F., Yuan, J., Tian, J., Zhang, Y., Shi, Z., Fan, J., He, Z.: A survey of visual transformers. IEEE Transactions on Neural Networks and Learning Systems, (2023)
19. Lu, D., Xie, Q., Wei, M., Gao, K., Xu, L., Li, J.: Transformers in 3d point clouds: A survey. arXiv preprint arXiv:2205.07417, (2022)
20. Mitchell, T. M.: The need for biases in learning generalizations. Readings in Machine Learning, (1980)

21. Mitchell, T. M., Mitchell, T. M.: Machine learning, vol. 1. McGraw-hill New York (1997)
22. Papillon, M., Sanborn, S., Hajij, M., Miolane, N.: Architectures of topological deep learning: A survey of message-passing topological neural networks. arXiv preprint arXiv:2304.10031, (2023)
23. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, vol. 32 (2019)
24. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, vol. 21, no. 140, pp. 1–67 (2020)
25. Rao, Y., Zhao, W., Tang, Y., Zhou, J., Lim, S.-N., Lu, J.: Hornet: efficient high-order spatial interactions with recursive gated convolutions. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22, Curran Associates Inc. (2022)
26. Tsai, Y.-H. H., Bai, S., Yamada, M., Morency, L.-P., Salakhutdinov, R.: Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 4344–4353. Association for Computational Linguistics, Hong Kong, China (Nov 2019) doi: 10.18653/v1/D19-1443 , https://aclanthology.org/D19-1443/
27. Vaswani, A.: Attention is all you need. arXiv preprint arXiv:1706.03762, (2017)
28. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks (2018), https://arxiv.org/abs/1710.10903