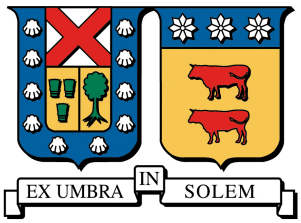
Universidad Técnica Federico Santa María
Departamento de Informática
Santiago - Valparaíso, Chile



Ajustando la similaridad intra-modal en un entrenamiento cross-modal basado en tripletas

Mario Carlos Mallea Ruz

Tesis para optar al Grado de
Magíster en Ciencias de la Ingeniería Informática

Profesor Guía: Ricardo Ñanculef A., Ph.D.
Profesor Co-Guía: Mauricio Araya, Ph.D.
Profesor Correferente Interno: Mauricio Solar, Ph.D.
Profesora Correferente Externo: María Soledad Pera, Ph.D.

29 de febrero de 2024

Abstract

Content-Based Image Retrieval (CBIR) is a technique that allows you to enter an image as a query, and retrieve the images that are most visually similar to the query in a database. A related technique is Cross-modal retrieval (CMR), which allows querying in one modality (e.g. text) and retrieving information in another modality (e.g. images). With the rapid growth of multimedia content, CBIR and CMR have become essential technologies for building information systems in various domains, such as: social networks, online retail, remote sensing, and medicine.

Cross-modal retrieval requires building a common latent space that captures and correlates information from different data modalities, usually images and texts. Cross-modal training based on the triplet loss with hard negative mining is a state-of-the-art technique to address this problem. This research shows that such approach is not always effective in handling intra-modal similarities. Specifically, we found that this method can lead to inconsistent similarity orderings in the latent space, where intra-modal pairs with unknown ground-truth similarity are ranked higher than cross-modal pairs representing the same concept. To address this problem, we propose two novel loss functions that leverage intra-modal similarity constraints available in a training triplet but not used by the original formulation. Additionally, this research explores the application of this framework to unsupervised image retrieval problems, where cross-modal training can provide the supervisory signals that are otherwise missing in the absence of category labels. Up to our knowledge, we are the first to evaluate cross-modal training for intra-modal retrieval without labels. We present comprehensive experiments on MS-COCO and Flickr30K, demonstrating the advantages and limitations of the proposed methods in cross-modal and intra-modal retrieval tasks in terms of performance and novelty measures.

Resumen

Content-Based Image Retrieval (CBIR) es un técnica que permite ingresar como consulta una imagen, y recuperar las imágenes visualmente más parecidas a la consulta en una base de datos. Una técnica relaciona es el Cross-modal retrieval (CMR), este permite consultar en una modalidad (ej: texto) y recuperar información en otra modalidad (ej: imágenes). Con el rápido crecimiento del contenido multimedia, CBIR y CMR se han convertido en tecnologías esenciales para construir sistemas de información en varios dominios, como: social networks, online retail, remote sensing , y medicine.

La recuperación cross-modal se basa en construir un espacio latente común, que capture y correlacione información proveniente de diferentes modalidades, usualmente imágenes y textos. Una técnica estado del arte para el diseño de este tipo de sistemas es el entrenamiento basado en tripletas con hard-negative mining. Esta investigación muestra que dicho entrenamiento no maneja siempre de manera efectica las similaridades entre elementos de la misma modalidad. Específicamente, en esta tesis se muestra que dicho método puede llevar a relaciones de similitud inconsistentes en el espacio latente construido, donde existen pares intra-modales que no corresponden a instancias de aprendizaje provista por los datos, pero que son priorizados por sobre los pares cross-modales que representan al mismo concepto y por ende, corresponden a instancias de aprendizaje recolectados en los datos. Para manejar dicho problema, se proponen dos nuevas funciones de pérdida que promueven restricciones sobre la similitud entre pares intra-modales, las cuales no son aprovechados en la formulación original. Adicionalmente, se exploró la aplicación de estos métodos a problemas de recuperación de imágenes no supervisado, donde el entrenamiento cross-modal puede proveer la señal de supervisión necesaria que se pierde en la ausencia de categorías. De acuerdo a la revisión bibliográfica, esta investigación es pionera en evaluar el entrenamiento cross-modal para recuperación intra-modal sin etiquetas. Se presentan experimentos en dos bien conocidos conjuntos de datos MS-COCO y Flickr30K, demostrando las ventajas y limitaciones de los métodos propuestos en escenarios de recuperación cross e intra modales, en términos de rendimiento y novedad de las recuperaciones.

Índice general

Índice de cuadros

Índice de figuras

Capítulo 1

Introducción

Esta tesis es motivada por una aplicación informática en el ámbito médico, en la cual se requiere el desarrollo de un sistema de búsqueda de imágenes (CBIR) que apoye el proceso de diagnóstico, con un sistema que permita distinguir automáticamente entre múltiples imágenes médicas que reflejan distintas condiciones, hallazgos o enfermedades [21]. Los datos de entrenamiento incluyen imágenes con sus respectivos reportes radiológicos, pero sin categorías explícitas que permita agrupar entre imágenes o reportes. En un escenario como este, esta investigación permitiría diseñar un sistema de recuperación de imágenes que se apalanca de la información textual para ajustar la representación de imágenes a través de un aprendizaje cross-modal.

La precisión de un mecanismo CBIR recae en gran medida en la representación de imágenes que es utilizada para calcular similitudes entre ellas. En la última década, el deep learning ha mejorado notablemente la extracción de características que producen algoritmos clásicos [5], resultando en representaciones de imágenes más discriminativas y robustas para tareas de CBIR [10]. Sin embargo, los métodos basados en deep learning requieren de un apropiado entrenamiento y, por tanto, el diseño de funciones objetivo o de pérdida que puedan guiar de manera apropiada el aprendizaje de la semántica asociada entre imágenes [12]. Una técnica estado del arte para el entrenamiento de modelos CBIR es el aprendizaje basado en tripletas [14, 5]. Un entrenamiento por tripletas consiste típicamente de una imagen ancla, una positiva (que es similar al ancla) y otra negativa (que es disimilar al ancla). La *triplet loss* es entonces usada para promover que la representación del ancla quede más cerca o sea más similar a la representación del positivo que la del negativo. En cross-modal retrieval [6], donde existen distintos tipos de datos, el aprendizaje de la representación consiste en construir un espacio latente común que capture y correlacione información desde distintas modalidades. La cross-modal *triplet loss* maneja este desafío tomando el ancla en una modalidad, y al positivo y negativo en la otra modalidad. De esta manera, este aprendizaje busca asegurar que la similitud entre pares cross-modales que representan el mismo concepto en diferentes modalidades sea mayor que la similitud entre pares cross-modales no observados en la data de entrenamiento. En la práctica, es bien conocido que este enfoque es gobernado por

los casos donde ocurre el mayor error. Estos elementos son conocidos como los *hard negatives* [6], los cuales son utilizados activamente durante el entrenamiento.

Particularmente, mi investigación muestra que el aprendizaje basado en tripletas cross-modales no maneja con precisión las similitudes entre elementos intra-modales, conduciendo a relaciones de similitud inconsistentes en el espacio latente. Este fenómeno puede introducir vecinos falsos en el espacio latente que perjudican el rendimiento de la recuperación. Investigaciones anteriores sobre este tema se han centrado en datos etiquetados, estudiando cómo la concentración dentro de cada clase y la dispersión entre clases contribuyen al rendimiento general del sistema [30, 7, 36, 43]. Sin embargo, sin datos etiquetados, no se pueden medir las similitudes intra e inter clases, ni se puede calcular fácilmente la precisión de las búsquedas dentro de una modalidad. Por lo tanto, comprender cómo ciertas propiedades del espacio afectan el rendimiento final del recuperador es aún más desafiante para el escenario de recuperación de imágenes sin etiquetas. Para abordar estas limitaciones, se proponen dos funciones de pérdida novedosas que utilizan restricciones de similitud intra-modalidad en un entrenamiento por tripletas cross-modales. A diferencia de la formulación clásica, las pérdidas propuestas promueven explícitamente que los pares cross-modales (que representan un mismo concepto) estén más cerca en el espacio latente en comparación con los pares intra-modales (los cuales no representan necesariamente un concepto similar). Además, se propone afrontar el problema de la evaluación de un sistema de recuperación intra-modalidad sin etiquetas, utilizando métricas de relevancia no binarias que aprovechen las modalidades adicionales para medir la relación semántica entre elementos [19, 18]. En particular, este enfoque hace posible la evaluación de sistemas de recuperación imagen-imagen, confiando en métodos que miden similitud entre grupos de textos que representan o describen los elementos.

Específicamente, los principales aportes son:

- Mostrar que la cross-modal *triplet loss* con hard negative mining puede conducir a órdenes de similitud inconsistentes en el espacio latente. En efecto, se muestra teórica y experimentalmente que los pares intra-modales no observados en los datos de entrenamiento, pueden estar más cerca que los pares cross-modales que representan el mismo concepto en diferentes modalidades.

- Proponer dos nuevas pérdidas basadas en tripletas que mejoran el uso de *hard negatives* a través de las relaciones de orden de similaridad cross e intra modales. Los resultados experimentales demuestran que las propuestas permiten la construcción de sistemas de recuperación cross e intra modales más precisos y novedosos.

- El diseño de una forma novedosa de evaluar tareas de recuperación de imagen a imagen sin etiquetar, a través de dos medidas de relevancia semántica que utilizan enfoques modernos de recuperación cross-modal. Además, se utiliza la información semántica para medir la novedad de las listas de recuperación.

También se utiliza el concepto de self-information para medir la sorpresa [44] general en los resultados obtenidos.

1.1. Hipótesis

La principal hipótesis de esta investigación es la siguiente:

Modificar el mecanismo estándar de entrenamiento vía tripletas para promover la consistencia de las relaciones de similaridad tanto intra-modales como cross-modales permite mejorar la precisión de un sistema de recuperación imagen-imagen en escenarios en que no se dispone de imágenes etiquetadas, pero sí de textos en lenguaje natural que uno o más humanos han usado para describirlas.

1.2. Objetivos

El objetivo general asociado de esta tesis es:

Implementar un modelo recuperador de imágenes por contenido, bajo un enfoque de entrenamiento cross-modal. La novedad recae en el diseño de una función de pérdida que permita un entrenamiento con técnicas estado del arte para recuperación cross-modal, pero que promueva la precisión de la tarea de recuperación imagen-imagen.

Los objetivos específicos asociados a esta tesis son:

1. Analizar e implementar las principales técnicas de recuperación cross-modal.
2. Formulación formal de una propuesta a partir del estudio descrito en el punto anterior. La propuesta se enfoca en la recuperación de imágenes a través del contenido visual y textual.
3. Evaluar el rendimiento e impacto de las técnicas propuesta sobre las recuperaciones.
4. Reportar los hallazgos de valor científico en 26th International Conference on Discovery Science, en Oporto, Portugal [17].

1.3. Estructura

El presente escrito comienza con una presentación formal del enfoque cross-modal usado para construir un sistema de recuperación (capítulo 2). Este es necesario para entender y formular las propuestas. Se sigue con una revisión sistemática de trabajo

relacionado con las técnicas utilizadas en esta investigación (capítulo 2.2). Posteriormente, se describe con ciertos elementos matemáticos el problema que motiva la formulación de los dos nuevos modelos que serán introducidos, también se presentan los detalles de las métricas adaptadas para evaluar los métodos presentados (capítulo 3). Seguido de ello, se presentan el diseño y metodología experimental que permiten entender cuantitativamente las ventajas y desventajas prácticas de los métodos propuestos, como sistemas recuperación intra y cross modales con respecto a métricas de precisión y novedad. A modo de análisis cualitativo de las características descritas, se presenta un ejemplo de recuperación (capítulo 4). Finalmente, se resume y se concluye sobre los puntos más relevantes que deja esta tesis, además de presentar los principales trabajos futuros posibles que se desprenden de los aportes de mi investigación (capítulo 5).

Capítulo 2

Marco Teórico

La recuperación de imágenes basada en contenido es un campo en constante evolución que busca mejorar la capacidad de los sistemas de búsqueda para encontrar imágenes relevantes dentro de grandes bases de datos. Uno de los enfoques relacionados en este contexto es el cross-modal retrieval, o recuperación cruzada de datos multimodales. Este enfoque se centra en la búsqueda y recuperación de información a través de múltiples modalidades, como imágenes, texto y otras formas de datos. Interesantemente, el diseño de este tipo de sistemas recae en entender y explotar las relaciones entre diferentes tipos de contenido.

En este marco teórico, se examina cómo las representaciones de datos y la similitud entre diferentes modalidades son esenciales para el diseño de sistemas de recuperación. Las cuestiones de evaluación y métricas relevantes para medir el rendimiento de estos sistemas serán discutidas en profundidad al final del capítulo 3.

2.1. Cross-Modal Retrieval

El modelo que se presenta en esta sección corresponde a VSE++ [6].

El objetivo de la tarea de recuperación cross-modal es aceptar una consulta en una modalidad y recuperar datos relevantes en la otra modalidad. Específicamente, consideremos N pares de imágenes y textos $P = \{(i_n, c_n)\}_{n=1}^{N}$ como data de entrenamiento, y sea p_{data} su distribución. Me referiré a estas tuplas como pares positivos, y llamaré pares negativos a aquellas tuplas que no pertenecen a P.

En este escenario, el aprendizaje por tripletas considera tanto la tarea de recuperación de imágenes a partir de textos (t2i), como la de recuperación de textos a partir de imágenes (i2t). El aprendizaje por tripletas formaliza la intuición que la similitud entre un par positivo (i_n, c_n) debería ser mayor que la similitud entre el texto de query c_n y alguna otra imagen (t2i), y análogamente para la tarea de i2t. La loss combina las dos tareas equitativamente:

$$\mathcal{L}(i_n, \bar{i_n}, c_n, \bar{c_n}) = \mathcal{L}_{i2t} + \mathcal{L}_{t2i}, \tag{2.1}$$

$$\mathcal{L}_{i2t} := [\alpha + s(i_n, \bar{c_n}) - s(i_n, c_n)]_+, \tag{2.2}$$

$$\mathcal{L}_{t2i} := [\alpha + s(\bar{i_n}, c_n) - s(i_n, c_n)]_+, \tag{2.3}$$

donde $[x]_+ = max(0, x)$ (tipo hinge loss [25]). Aquí α es un hiperparámetro conocido como margen. La similaridad s se calcula como el producto punto $s(i, c) = f_v(i)^t f_t(c)$ entre las representaciones normalizadas [6] f_v y f_t asignada a imágenes y textos, respectivamente. El encoder visual es $f_v(i) = E^V g_i$ y el encoder textual es $f_t(c) = E^T t_c$, donde $g_i \in \mathbb{R}^{n_V}$ es el vector de características asociado con la imagen i y $t_c \in \mathbb{R}^{n_T}$ para el texto. Si utilizamos redes neuronales preentrenadas para obtener estos vectores de características, entonces los parámetros entrenables θ son las matrices de proyección $E^V_\phi \in \mathbb{R}^{k \times n_V}$ y $E^T_\phi \in \mathbb{R}^{k \times n_T}$, con k el hiperparámetro que define la dimensión del espacio latente.

El ajuste de los parámetros entrenables se logra a través del algoritmo del gradiente descendiente estocástico aplicado al siguiente problema de optimización:

$$\min_{\theta} \mathop{\mathbb{E}}_{\substack{(i_n, c_n) \sim p_{data} \\ (\bar{i_n}, \bar{c_n}) \sim \bar{p}_{data}}} \mathcal{L}(i_n, \bar{i_n}, c_n, \bar{c_n}). \tag{2.4}$$

En la práctica, el valor esperado es aproximado muestreando desde la data de entrenamiento. Un enfoque simple es considerar también el muestreo aleatorio de los negativos. Sin embargo, es bien conocido que una mejor técnica es la de hard negative mining, donde los negativos son escogidos dinámicamente de acuerdo al estado del modelo actual [6]. Ósea, $\bar{p}_{data} = p^{max}_{data}$, lo que significa que se toman como negativos los casos que son más problemáticos para cada tarea y en cada paso del algoritmo de optimización:

$$\bar{i_n} = \mathop{argmax}_{x \neq i_n} s(x, c_n), \text{ y } \quad \bar{c_n} = \mathop{argmax}_{x \neq c_n} s(i_n, x) \tag{2.5}$$

La Fig 2.1 es una ilustración del proceso de entrenamiento, dos componentes del modelo son conjuntamente entrenados sobre pares en la modalidad visual y textual, con el objetivo de maximizar la similitud entre los pares de entrenamiento. Basado en el par positivo, los hard negatives son recolectados, y luego procesados por modelos con parámetros compartidos en cada modalidad.

2.2. Trabajo Relacionado

El aprendizaje basado en tripletas, y particular con hard negative mining, son técnicas ampliamente utilizadas en diferentes áreas como: deep learning metric, [12], deep face recognition [34], person re-identification [41], recommender systems [2], content-based image retrieval [5], y cross-modal retrieval [5], entre otros.

En el área de deep metric learning (DML) [39, 12], la siguiente investigación [27] adapta el aprendizaje por tripletas, de manera que en vez de iterar el proceso de

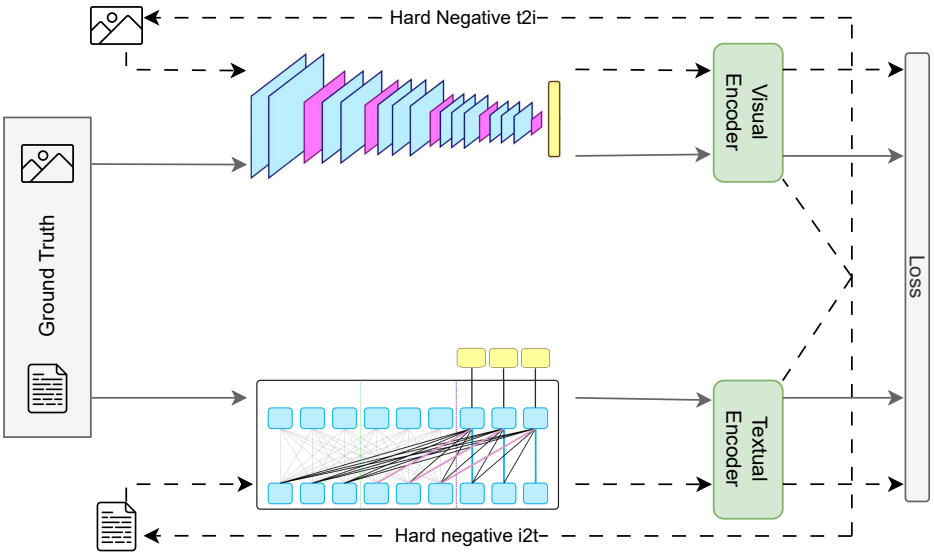


Figura 2.1: Proceso de entrenamiento cross-modal. La línea entre cortada guía el proceso del cálculo de los negativos, mientras que la línea continua guía el flujo de entrenamiento.

optimización por tripletas individuales, se propone una nueva formulación de pérdida que explota o aprovecha todas las relaciones de pares posibles dentro de un minibatch de entrenamiento. Sin embargo, este enfoque confía en la existencia de imágenes con clases etiquetadas para construir dichas muestras de entrenamiento. Otro trabajo que adapta la pérdida por tripletas en DML es [4], este propone una linealización de la pérdida basada en tripletas, que garantiza complejidad lineal al considerar centroides fijos por cada clase. Idealmente, dichos centroides serán representativos de las clases en el espacio latente, lo cual permite ajustar las relaciones de orden de similitud entre elementos a través de los centroides. En principio no es trivial la adaptación de dicho método para un escenario sin etiquetas. Otro aspecto importante del aprendizaje por tripletas es el efecto del margen, esto es, la diferencia mínima permitida entre la similitud de elementos de la misma clase y la similitud entre elemento de diferentes clases. Con el propósito de estudiar el efecto del margen en la selección de los hard negatives, esdecir, los casos más problemáticos durante el entrenamiento, [7] propone calcular dinámicamente el margen usando la cota utilizada para unir dos clases en la construcción de una jerarquía a partir de las distancias promedio por clase, con el propósito de estimar la concentración intra-clase y la dispersión inter-clase. Adicionalmente, otra investigación como [35] define un margen adaptativo que es calculado como una función no lineal y no entrenable. Esta función recive como input el promedio de las distancias entre pares positivos (misma clase) y pares negativos, los cuales son generados a partir de un modelamiento adversarial. Por lo que la función a optimizar considera tanto la componente de aprendizaje por tripletas como el adversarial. Nuevamente, su objetivo es compactar las distancias

intra-clase y separar las inter-clase. De estos últimos dos trabajos, es posible inferir que la componente del margen tiene una relación directa con las características del espacio latente que permiten discriminar entre clases, por lo que es útil definir el margen a partir de alguna función de las distancias entre clases, sin embargo, en un escenario sin etiquetas, no es evidente cómo medir el impacto que podría tener adaptar el margen en un aprendizaje por tripletas.

En particular, cuando se utiliza un aprendizaje guiado por una medida de similaridad [37], y en un escenario de aprendizaje cross-modal (pero con data etiquetada), la investigación [38], propone un margen adaptativo que considera la información intra-modal, similar a uno de mis modelos propuestos. Por ejemplo, dado un texto, el margen es definido como una combinación convexa de la distancia entre imágenes en el espacio de características y en el espacio de la representación one-hot de etiquetas. Cabe destacar que también son presentadas algunos argumentos a favor de considerar una auto calibración del margen durante el entrenamiento, los cuales permiten inspirar algunos de los argumentos utilizados para caracterizar uno de mis métodos propuestos.

En la tarea de person re-identification [41], [36] introduce una quadruplet loss que extiende la triplet loss incorporando un segundo negativo, el cual es recolectado de manera que corresponda a una clase diferente al negativo clásico. Con ello, se busca promover que la distancia entre ambos negativos sea mayor que la distancia del par positivo. Lo cual tiene una cierta semejanza a una de las restricciones que incorporo en uno de los modelos propuestos. Así, implícitamente, se promueve que la mínima distancia inter-clase sea mayor que la distancia máxima intra-clase. Por otro lado, la investigación [43] también optimiza las distancias intra/inter clases y reduce el costo asociado con el cálculo del hard negative. Esto lo logra aprendiendo centroides de las clases (tal como [4]) y usándolos como anclas para calcular los hard negative.

En la tarea de image retrieval con descriptores globales de la imagen. La terminología de descriptor global, es usado para diferenciarse de enfoques de representación secuenciales o de regiones de interés [24] de la imagen. La investigación [31] extiende la triplet loss con hard-negative mining, ya que considera añadir un término de regularización [30]. Este término fuerza que, en promedio (por minibatch), la distancia entre consultas sea aproximadamente la misma que la distancia entre los positivos seleccionados para dichas consultas. Los autores lo destacan como un método de regularización de segundo orden del espacio latente, ya que por sí mismo no es capaz de construir un espacio latente discriminativo entre clases. El beneficio de dicho método es medido a través de cantidades de dispersión-concentración por clase de la distribución de vectores en el espacio latente construido.

Todas estas investigaciones previas muestran que es posible mejorar el aprendizaje por tripletas en variadas aplicaciones. Estas mejoras se realizan a través de diversos enfoques, los cuales se vinculan con diferentes propiedades del espacio latente construido. Por tanto, los métodos propuestos en este trabajo atacan un método relevante actualmente y que está en constante evolución.

Específicamente en cross-modal retrieval, VSE++ [6] es el modelo que demostró la efectividad de la triplet loss con hard negative mining, este constituye el principal baseline en mi investigación, este modelo será explicado en detalle en el marco teórico, capítulo 2. El estado del arte para el marco experimental de VSE++, resulta de un mejor diseño de la arquitectura neuronal y no de modificar la técnica de entrenamiento [18, 19]. Estos modelos superiores utilizan representaciones finas o granulares del contenido (como regiones de interés) y son procesadas por arquitecturas transformers [33] para su alineamiento contextual. Es decir, la similitud entre elementos es calculado a partir de una matriz de similitud y no como el producto interno de una sola representación por contenido. Además, estos trabajos proponen una métrica que permite evaluar de manera más fina el éxito de la recuperación cross-modal. Esta métrica será utilizada activamente, por lo que será explicada en detalle en el capítulo 3.

Otro aporte relevante de esta investigación se relaciona con un análisis teórico que sirve de motivación a los modelos propuestos. Particularmente, se trata de un problema vinculado al no manejo de las relaciones intra-modales en la cross-modal triplet loss. Este problema ya ha sido previamente identificado en la literatura, pero en otros contextos, como por ejemplo de pre entrenamiento de modelos de lenguaje-visión con técnicas de aprendizaje no supervisado [40], y también en un contexto de recuperación uni-modal de texto [23]. En ambos, y a diferencia del problema que se describe en este trabajo, se utilizan una lista de negativos en batch procesados por una formulación de función de pérdida tipo log-likelihood, en vez de la utilizada por los modelos estado del arte en cross-modal retrieval, que es una basada en tripletas (individuales) aplicando una hinge loss con hard negatives mining. Por lo tanto, hay que considerar que, como es común en las técnicas de aprendizaje no supervisado, en [40] la relación intra-modal que se promueve es entre un elemento en una modalidad y una aumentación del mismo, es decir, un par compuesto por una imagen y su rotación u otra transformación visual. Esto es un diferenciador al problema expuesto en este trabajo, en el que la relación intra-modal es de los componentes que se utilizan en la cross-modal triplet loss, es decir, una imagen de consulta y su hard-negative.

En resumen, investigaciones previas se han enfocado en escenarios de data etiquetada, donde es posible estudiar como la concentración intra-clase y la dispersión inter-clase contribuyen a la performance de recuperación. Sin embargo, sin clases anotadas, las similitudes intra/inter clases no pueden ser medidas, y más aún, la precisión de un sistema de recuperación intra-modal no puede ser medida trivialmente. Por lo tanto, entender de qué manera ciertas propiedades del espacio afectan a la performance final del recuperador es aún más desafiante para un escenario de unlabeled image retrieval. Con base a mi revisión de la literatura, soy el primero que en un escenario cross-modal sin etiquetas, extienda o modifique la triplet loss con hard negative mining con el objetivo de reparar las relaciones de orden de similitud intra-modales. Además de evaluar la tarea de image retrieval a través de un entrenamiento cross-modal (sin necesidad de recurrir a etiquetas).

Capítulo 3

Modelos propuestos

Propongo el diseño de un sistema de recuperación de imágenes a través de un entrenamiento cross-modal. Sin embargo, la formulación descrita en el capítulo 2, presenta ciertos problemas para cumplir este objetivo. Una vez que se identifican dichos problemas, propondré dos nuevas formulaciones que también se basan en el aprendizaje por tripletas con hard negative mining, pero que, a diferencia de la formulación clásica, integrarán al aprendizaje las relaciones de orden de similitud entre elementos de la misma modalidad. Cabe destacar que estas nuevas técnicas propuestas no incrementan considerablemente la complejidad computacional del entrenamiento clásico, pues no modifican la técnica de muestreo de los elementos durante el entrenamiento. Adicionalmente, propongo evaluar la recuperación intra-modal usando métricas de relevancia no binaria que aprovechan las modalidades adicionales para medir la relación semántica entre elementos [19, 18]. Este enfoque permite evaluar sistemas de recuperación intra-modal sin necesidad de clases etiquetadas. Además, para evaluar el sistema más allá de su performance de recuperación, consideré la evaluación con respecto a la novedad de las recuperaciones.

3.1. Definición del Problema

Una de las desventajas de la cross-modal triplet loss es que solo considera relaciones de similiaridad cross-modales, y, por tanto, ignora el impacto de la similitud entre elementos intra-modales. Sin pérdida de generalidad, vamos a considerar una tarea de recuperación t2i para ilustrar las limitaciones de la formulación clásica. Dado una tripleta de entrenamiento $(i, c, \bar{i})$, la triplet loss tipo hinge loss (Eqn. 2.3) promueve la restricción $s(i, c) > s(\bar{i}, c) + \alpha$, la cual asegura que la imagen i será más similar que la imagen hard negative $\bar{i}$ con respecto a la consulta c (muestreado de los datos como el texto asociado a la imagen i). Sin embargo, esta restricción no impone condiciones de la relación de similitud entre i y $\bar{i}$. Como se ilustra en la Fig 3.1, aunque el texto de consulta c es más similar a su imagen i que a la imagen hard negative $\bar{i}$ (satisfaciendo el criterio impuesto), ante un insuficiente margen es posible

que $\bar{i}$ sea más similar a i en comparación a c. En otras palabras, existen $(i, c, \bar{i})$ tales que:

$$s(i, \bar{i}) > s(i, c) > s(\bar{i}, c) + \alpha \ . \tag{3.1}$$

Identifico que esta situación puede ser problemática por varias razones:

1. Mientras el par (i, c) se sabe que representa el mismo concepto, el par $(i, \bar{i})$ podría no corresponder a conceptos similares. Sin una supervisión explícita (como clases etiquetadas), es desafiante diferenciar entre relaciones de similaridad válidas. Entonces, como solo (i, c) es observado en los datos, deberíamos esperar que $s(i, c)$ sea mayor que $s(i, \bar{i})$.

2. Sea $c_{\bar{i}}$ un texto que corresponde a $\bar{i}$. Si $s(i, \bar{i})$ y $s(i, c)$ son altos (como se esperaría que suceda), c puede ser priorizado o rankeado mejor que $c_{\bar{i}}$, cuando $\bar{i}$ sea presentado como una consulta, resultando en un error para la tarea de recuperación i2t. Si bien la triplet loss podría manejar este problema a través de la restricción i2t, corregir dicho error podría requerir pasos de entrenamiento adicionales, ralentizando el entrenamiento.

3. Sin formas adicionales de supervisión, esperamos que un alto valor de $s(i, \bar{i})$ (representación de imágenes similares), suceda si y solo si, $s(c, c_{\bar{i}})$ también es alto (similar representación de textos). Sin embargo, si se tiene Eqn. 3.1, es posible que $s(c, c_{\bar{i}})$ sea pequeño, pero $s(i, \bar{i})$ grande (respecto uno de otro). Pues si $\bar{i}$ es muy similar a $c_{\bar{i}}$, podría pasar que $s(c, c_{\bar{i}}) \approx s(c, \bar{i})$, resultando en la siguiente desigualdad: $s(i, \bar{i}) > s(i, c) > s(c, c_{\bar{i}}) + \alpha$, lo cual muestra que la relación en la modalidad visual $s(i, \bar{i})$ es inconsistente con la relación en la modalidad textual $s(c, c_{\bar{i}})$, ya que la similitud entre elementos que no representan un mismo concepto en una modalidad resultan más similar que el par positivo, pero en la otra modalidad resulta menos similar que el par positivo.

En efecto, para representaciones latentes normalizadas, se puede entender que tan probable es que existan tripletas que cumplan el problema identificado anteriormente, al resolver Eqn. 3.1. En la Fig. 3.2 se presenta una solución numérica, concretamente, para ciertos ángulos positivos p (lo que se ven como cortes transversales en la figura), se considera una malla discreta de los otros dos ejes (los ángulos de un par negativo n, y entre una consulta y un hard negative en la misma modalidad im). Entonces se destacan aquellos que satisfacen el sistema de inecuaciones, y también se considera la desigualdad triangular $im \geq n - p$ [26] para evitar tripletas que no son posibles. Esto se resolvió en dos casos, para un margen de 0,4 y sin margen. Una de las conclusiones, es que la existencia de tales tripletas es naturalmente influenciado por la magnitud del margen. Un margen mayor ayuda a reducir la probabilidad de seleccionar estas tripletas, pues nótese el volumen más pequeño en la Fig. 3.2.b, pero es necesario

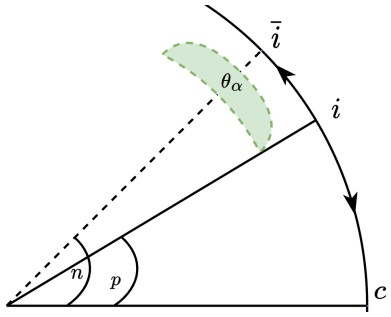
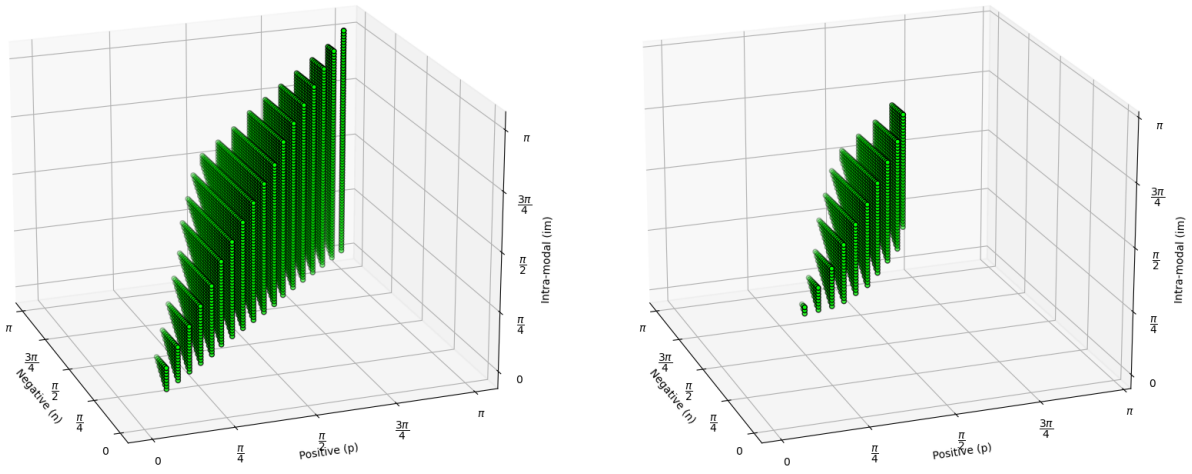


Figura 3.1: Ejemplo de una distribución inconveniente de vectores para la tarea de t2i.



(a) $cos(im) \geq cos(p) \geq cos(n)$.

(b) $cos(im) \geq cos(p) \geq cos(n) + 0{,}4$.

Figura 3.2: Muestra de la existencia empírica de tripletas problemáticas.

considerar que no es suficiente para eliminar el problema. Entonces, en el modelo clásico estas tripletas problemáticas son indirectamente controladas por el margen. Sin embargo, un margen fijo es una estrategia subóptima, porque podría no adaptarse a diferentes configuraciones de tripletas, permitiendo tripletas problemáticas y, por tanto, no favoreciendo la capacidad discriminativa y de generalización del modelo.

3.2. Modelos propuestos

3.2.1. Full hard negative method (F-HN)

Propongo hacer un mejor uso del hard negative extendiendo la loss para considerar todas las restricciones de similitud que pueden derivarse de un entrenamiento por tripletas. Por lo tanto, se añaden tres nuevos términos a la formulación clásica:

$$\mathcal{L}(i_n, \bar{i_n}, c_n, \bar{c_n}) = \mathcal{L}_{i2t} + \mathcal{L}_{t2i} + \mathcal{L}_{vc} + \mathcal{L}_{tc} + \mathcal{L}_{sc}. \tag{3.2}$$

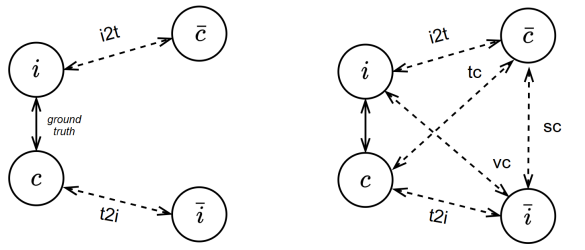
Los nuevos componentes de la loss se denominarán *restricción visual* (vc), *restricción textual* (tc) y *restricción estructural* (sc). Se definen de la siguiente manera:

$$\mathcal{L}_{vc} := [\alpha + s(i_n, \bar{i_n}) - s(i_n, c_n)]_+, \tag{3.3}$$

$$\mathcal{L}_{tc} := [\alpha + s(c_n, \bar{c_n}) - s(i_n, c_n)]_+, \tag{3.4}$$

$$\mathcal{L}_{sc} := (\mathbb{1}_{(\bar{i_n}, \bar{c_n}) \notin P})[\alpha + s(\bar{i_n}, \bar{c_n}) - s(i_n, c_n)]_+. \tag{3.5}$$

Aquí $\mathcal{L}_{vc}$ y $\mathcal{L}_{tc}$ tienen como objetivo garantizar que las similitudes intra-modales $s(i_n, \bar{i_n})$ y $s(c_n, \bar{c_n})$ son inferiores a los de los pares positivos cross-modales. Mientras tanto, L_{sc} actúa solo en los casos en que los hard negatives para cada tarea cross-modal no se corresponden entre sí y, por lo tanto, no deberían ser más similares que el par positivo o ground truth. Se puede ver la diferencia con el modelo clásico de tripletas en la Fig. 3.3, notar como F-HN explota todas las posibles relaciones de similitud entre los mismos elementos que ya se consideraban en la formulación clásica.



(a) Vanilla triplet loss.

(b) Proposed triplet loss.

Figura 3.3: Representación de relaciones de orden promovidas por el modelo clásico HN y el propuesto F-HN.

3.2.2. Intra-modal margin hard negative control method (M-HN)

Propongo el uso de un margen adaptativo para cada tripleta de entrenamiento, este margen permite la selección de muestras informativas localmente, capturando estructuras de similitud locales en el espacio latente y haciendo el proceso de entrenamiento más eficiente.

Para una tripleta dada $(i_n, c_n, \bar{c_n})$, se determinan los valores de margen en función de las similitudes intra-modales. En el caso de la componente de loss i2t, se busca asegurar que la restricción cross-modal se mantenga con un margen que sea al menos la similitud intra-modal de consulta $s(i_n, \bar{i_n})$. Por supuesto, el análisis para t2i es simétrico. Esta forma de modelar lleva a las siguientes definiciones:

$$\mathcal{L}(i_n, \bar{i_n}, c_n, \bar{c_n}) = \mathcal{L}^*_{i2t} + \mathcal{L}^*_{t2i}, \tag{3.6}$$

$$\mathcal{L}^*_{i2t} := [s(i_n, \bar{i_n}) + s(i_n, \bar{c_n}) - s(i_n, c_n)]_+, \tag{3.7}$$

$$\mathcal{L}^*_{t2i} := [s(c_n, \bar{c_n}) + s(\bar{i_n}, c_n) - s(i_n, c_n)]_+. \tag{3.8}$$

En comparación con la clásica cross-modal triplet loss, la formulación propuesta introduce una restricción i2t más estricta para las tripletas $(i_n, c_n, \bar{c_n})$ con representaciones de imágenes similares y una restricción i2t más flexible para las tripletas con representaciones de imágenes disimilares. Pues al actuar la similaridad intra-modal como margen, se puede interpretar que ante una consulta más confusa o difícil de discriminar en el espacio latente, se impondrá un mayor margen de separabilidad para contribuir en el ajuste de un espacio latente más discriminativo. Debido a que un mayor margen implica corregir más tripletas vía grandientes, o equivalentemente, ser más selectivo al determinar una tripleta con loss mínima (de valor 0). Este es un efecto de control automático. Además, este enfoque promueve una mayor coherencia entre las modalidades visual y textual, lo que lleva a un aumento o disminución de la similitud tanto cross-modal como intra-modal. Esta coherencia es beneficiosa para la recuperación cross-modal, así como, para escenarios de recuperación intra-modal, donde la única señal de supervisión disponible para aprender una representación se obtiene de la otra modalidad. Finalmente, se debe tener en cuenta que esta formulación elimina el margen como hiperparámetro, lo que lo hace un método más fácil de implementar y menos propenso al sobre ajuste.

3.3. Evaluación propuesta

La tarea de recuperación cross-modal se evalúa comúnmente utilizando el recall entre los primeros 1, 5 y 10 resultados recuperados, es decir, si se recupera o no (binario) la imagen ante la consulta de uno de sus textos, y viceversa. Además de esta forma de evaluación clásica, se amplió la evaluación utilizando una métrica de relevancia no binaria propuesta recientemente en [19, 18], que se nombra como

nDCG semántico (para un top de $K = 25$). A continuación se explica cómo se amplía esta métrica para evaluar la recuperación intra-modal. También se propone ampliar la evaluación considerando métricas de novedad, llamadas novelty-biased nDCG y novedad vía self-information. La medición de novedad de recuperación es relevante desde la perspectiva del usuario, por ejemplo, en el escenario médico en que los usuarios usan activamente el sistema, por lo tanto, requieren que el sistema no sea repetitivo. Las métricas de novedad que se introducen miden dicho aspecto, mientras que este aspecto es débilmente considerado por las métricas de performance de recuperación.

3.3.1. nDCG semántico

Por ejemplo, para la tarea t2i, dado un texto, la métrica nDCG tradicional ocupa una función de relevancia binaria, ya que penaliza la posición de recuperación de la imagen solo para la imagen correspondiente. Notar que en un escenario alimentado solo por datos en forma de pares, no existe de manera directa otra manera de evaluar. En este trabajo, promuevo utilizar una medida de relevancia no binaria, donde cada imagen que se recupera tiene una puntuación de relevancia que se obtiene comparando la similitud entre la consulta de texto y el grupo de textos que describen la imagen recuperada.

Formalmente, para una consulta q, el nDCG@K es definido como $\text{nDCG}_K = \text{DCG}_K/\text{IDCG}_K$, donde $\text{DCG}_K = \sum_{v=1}^{K} \frac{\text{rel}(q,v)}{\log_2(v+1)}$, y IDCG_K es una constante de normalización que es independiente del modelo y que hace que $\text{nDCG}_K = 1$ para una lista ideal de recuperaciones, en el sentido tener el mejor ranking posible en función de la similitud que aquí se mide.

Además, $rel(q, v)$ es una función de relevancia que se calcula dependiendo de la modalidad:

- Recuperación de imagen. En las tareas de t2i e i2i, se calcula $\text{rel}(q, v) = \tau\left(\bar{C}_v, C_q\right)$, donde τ ya será definido. C_q es el caption o texto de consulta y $\bar{C}_v$ es el conjunto de todos los captions o textos asociados a la imagen recuperada I_v. Entonces, en el caso de la tarea i2i, el top de recuperaciones visuales, se evalúa a través del promedio de la métrica definida para cada uno de los diferentes captions o textos que corresponden a la consulta. Esto debido a que la relevancia entre el grupo de textos que corresponden a la imagen recuperada y cada uno de los textos asociados a la imagen de consulta es diferente [19].

- Recuperación de texto. Para las tareas de i2t y t2t, se calcula $\text{rel}(q, v) = \tau\left(\bar{C}_q, C_v\right)$, donde $\bar{C}_q$ es el conjunto de captions asociados a la imagen de consulta I_q. En caso de evaluar la tarea t2t, se construye el ranking o top con base a la similaridad entre las representaciones de los textos aprendidas con el modelo. Y se evalúa utilizando la función de relevancia asociada a la imagen que corresponde a la consulta, esto se debe a que para la tarea de i2t, [19]

calcula la relevancia del grupo de textos asociados a la imagen de consulta hacia los captions recuperados, y no de manera específica para cada caption. Esto implica que los resultados de recuperación texto a texto se evalúan por su rendimiento conjunto entre los captions que describen a la misma imagen que describe la consulta.

Siguiendo el estado del arte en recuperación cross-modal [18, 19] se utilizó Rouge-L [15] y Spice [22] como funciones τ de similitud entre oraciones. Estas dos funciones capturan diferentes aspectos de las oraciones.

- Rouge-L: A partir de la evaluación de la tarea de resumen automático, es decir, la comparación entre resúmenes generados por IA en comparación al trabajo realizado por humanos, Rouge-L proporciona información sintáctica, pues opera calculando la subsucesión de palabras comunes más largas entre frases, lo que permite no tener que predefinir una longitud de n-gramas ni necesitar de palabras exclusivamente en orden consecutivo, sino que captura palabras en secuencia.
- Spice: A partir de la evaluación para la tarea de generar automáticamente descripciones de imágenes, Spice considera características semánticas en las frases. Esta explota técnicas que utilizan modelos pre entrenados en grandes corpus de lenguaje con el objetivo de crear grafos de escena a través de un análisis sintáctico-semántico, por ejemplo, estas técnicas consultan la lematización, clasificación gramatical y sinónimos de wordnet [20]. De esta manera, la métrica consiste en un F-score sobre las tuplas de proposiciones semánticas en los grafos de escena, en palabras simples, contando repeticiones de relaciones entre objetos, atributos y relaciones gramaticales.

Ahora bien, por razones de eficiencia, estas funciones de relevancia entre frases no se calculan entre todas las disponibles. Si no que exclusivamente para las consultas de cada fase de evaluación (ya sea validación o test).

3.3.2. Novelty-biased nDCG

Propongo una extensión de $\alpha-nDCG@K$ [3], de manera de utilizar las funciones de relevancia no binarias (el trabajo original es solo binario). Para una consulta q, la métrica se define como:

$$\frac{1}{IDCG}\sum_{v=1}^{K}\frac{rel(q,v)(1-\alpha)^{r(q,v-1)}}{log_2(v+1)}, \text{ donde } r(q,v-1)=\sum_{z=1}^{v-1}rel(q,z)\,,$$

La constante de normalización $IDCG$ se estima considerando los resultados más relevantes primero para cada consulta. Se considera $\alpha(= 0{,}5)$ como una constante que penaliza largas repeticiones de resultados relevantes en favor de la novedad en función de la relevancia acumulada r hasta el momento. La función de relevancia rel se aplica de acuerdo a la tarea, tal como se explicó para el $nDCG$ semántico.

3.3.3. Novedad via self-information

También se propone la adaptación de una métrica que se usa en el área de sistemas recomendadores [44] para medir cuáles recomendaciones resultan novedosos, en el sentido, de no ser esperados o usuales. El concepto de *self-information* puede ser interpretado como un concepto de sorpresa de una recuperación, este se define como el inverso de la probabilidad a priori de seleccionar dicha recuperación, la métrica se define como:

$$\frac{1}{K}\sum_{v=1}^{K}\log_2\left(\frac{n_q}{count(v)}\right),$$

donde $count(v)$ cuenta el número de veces que v fue recuperado durante las n_q consultas de evaluación. Notar entonces que no existe novedad al recuperar un elemento que ha sido recuperado por todas las consultas, y entonces esta métrica premia la diversidad de recuperaciones.

En resumen de ambas maneras propuestas de medir la novedad de un sistema, con el concepto de self-information se recompensan resultados que son globalmente infrecuentes a lo largo de todas las listas de recuperaciones, mientras que, Novelty-biased nDCG recomenpensa resultados que quiebran la monotonicidad dentro de una lista individual de recuperaciones en función de la posición y relevancia de las recuperaciones.

Capítulo 4

Experimentos

Los métodos propuestos serán evaluados y comparados con los principales baselines que permitan concluir acerca de los beneficios impuestos en las formulaciones, esto en términos de precisión y novedad para tareas de recuperación cross e intra modales en un escenario en los cuales solo se tienen pares de imágenes y textos.

4.1. Datasets

Aunque propongo extender la evaluación, también utilizo marcos experimentales ampliamente utilizados en la investigación de recuperación cross-modal [18, 6, 9, 19]. En particular, se utilizan conjuntos de datos populares en el área: MS-COCO [16] y Flickr30k [42]. En el cuadro 4.1 se presentan las estadísticas con respecto al número de imágenes. Cada conjunto de datos consta de un conjunto de imágenes que cuentan con cinco captions o descripciones escritas por humanos. Se usaron particiones ampliamente usadas en la literatura para definir los conjuntos de la validación y pruebas (introducidas por [11]).

Cuadro 4.1: Estadística de los conjustos de datos.

	Train	Validation	Test	Total
Flickr30k	29,000	1,014	1,000	31,014
MS-COCO	113,287	5,000	5,000	123,287

4.2. Modelos y detalles de implementación

Se comparan los métodos propuestos F-HN y M-HN con los siguientes baselines:

- RN: Este método utiliza la clásica cross-modal triplet loss seleccionando aleatoriamente el negativo dentro del mini-batch [6]. RN es equivalente a VSE0 [6] pero, como se detalla a continuación, se utilizan arquitecturas más recientes para extraer representaciones de las imágenes y los textos.

- HN: Este método utiliza la clásica cross-modal triplet loss con hard negative mining, lo cual es equivalente a VSE++ [6] pero con representaciones de imagen y texto actualizadas.

- TERAN: Este es un método que también utiliza la clásica cross-modal triplet loss con hard negative mining, pero codificando el contenido a través de una arquitectura tipo Transformer, lo cual permite calcular la similitud entre una imagen y un texto por medio de un mecanismo atencional entre representaciones de granulares o desenredadas del contenido. Este modelo es estado del arte para las tareas cross-modales [18].

- ZS: Este un método tipo Zero-Shot que propongo, es decir, sin un entrenamiento específico. Para ello se utilizan las representaciones preentrenadas del contenido visual y textual. Dado que naturalmente dichas representaciones son de una dimensionalidad diferente, para hacer posible la recuperación cross-modal, se realiza reducción de dimensionalidad del contenido que se representa en una mayor dimensionalidad hacia la dimensión menor. Se experimentó ajustando TSNE vía la perplexity [32], pero los mejores resultados se obtuvieron con PCA [8].

Para una justa, rápida y simple experimentación, se utilizaron redes neuronales previamente entrenadas para la extracción de características. Por tanto, solo entreno la proyección de las representaciones visuales y textuales hacia el espacio latente común. Sé utilizan modelos de última generación para ambas modalidades, lo cual implica una actualización de los resultados de VSE [6]. Para las imágenes se utilizó el modelo Efficient Net V2L ($n_V = 1280$) [29], y para el texto se utilizó el modelo MPNET ($n_T = 768$) [28]. Las características visuales se calcularon en toda la imagen, sin técnicas de aumentación como extracción de recortes aleatorios [6].

La metodología comienza seleccionando los mejores hiperparámetros de acuerdo con la suma de Recall en las tareas de recuperación cross-modal [6]. Para Flickr30K, se seleccionó aleatoriamente la mitad de los datos de entrenamiento, y lo evalué en el conjunto de validación completo. Para compararse con la mejor versión posible del método HN, y así poner a prueba la hipótesis en un escenario difícil o complejo, se ajustó el margen entre $0,2$-$0,4$ y la dimensionalidad del espacio latente en $[768, 896, 1024]$. Los mejores resultados se obtuvieron para un margen de 0,4 y una dimensionalidad de 1024[1]. Estos valores fueron transferidos a los otros modelos. Se exploró modificar el margen en los nuevos componentes de F-HN considerando la grilla $[1e-3, 1e-3, 0{,}2, 0{,}4, 0{,}6]$ [2]. El mejor resultado se obtuvo con el mismo margen de 0,4 seleccionado para la clásica cross-modal triplet loss. Para MS-COCO, se adaptaron los valores de hiperparámetros comúnmente utilizados en la literatura: un margen de 0,2 y una dimensionalidad igual a 1024, que también provienen de una

[1] Con resultados de $[351{,}3 - 350{,}3, 351{,}0 - 349{,}3, 350{,}4 - 351{,}4]$, respectivamente.
[2] Con resultados de $[367{,}0, 368{,}1, 375{,}9, 377{,}5, 370{,}4]$, respectivamente.

optimización de hiperparámetros para el modelo HN [6]. Se siguió la recomendación de mantener iguales los márgenes de las nuevas restricciones en F-HN, y por razones de rapidez se trabajó con un mini batch de tamaño 512. Los modelos fueron entrenados para 20 épocas con el optimizador Adam [13] usando una tasa de aprendizaje de 0,0002 para las primeras 15 épocas y 0,00002 para las épocas restantes [6]. En conclusión, esta metodología implica comparar los modelos propuestos en un escenario ajustado para el clásico modelo HN, eventualmente se podrían obtener incluso mejores resultados si se tunea el margen y la dimensionalidad del espacio latente específicamente para cada modelo propuesto (F-HN y M-HN).

4.3. Resultados y discusión

Las cuadros 4.2, 4.3, 4.4, y 4.5 presentan los resultados [3] para las tareas de recuperación i2t, t2i, i2i y t2t, respectivamente.

4.3.1. Rendimiento de recuperación cross-modal

Los resultados de ZS demuestran que existe un desalineamiento notorio entre las distribuciones preentrenadas, lo que lleva a pobres resultados de recuperación cross-modal. Por lo tanto, el entrenamiento basado en tripletas tiene un gran efecto al ajustar las proyecciones lineales o reductores de dimensionalidad, por sobre lo que se puede obtener con un enfoque clásico como PCA.

Para la tarea i2t, se puede observar que HN supera a los otros algoritmos en términos de Recall para el top 1 y 5, pero F-HN es superior para el top 10 en Flickr30k. En general, F-HN se ubica como el segundo mejor algoritmo. Vale la pena mencionar que los resultados para RN y HN superan los resultados disponibles públicamente para VSE0 (ResNet-GRU) y VSE++ (ResNet-GRU) [6], respectivamente. Estos hallazgos proporcionan evidencia del rendimiento superior de las arquitecturas neuronales utilizadas para obtener representaciones visuales y texuales (Efficient Net-MPNET). No obstante, es importante reconocer que el rendimiento de todos los métodos se pueden mejorar aún más con un entrenamiento end-to-end, es decir, si se integra el ajuste o fine-tunning de los extractores de características neuronales para la tarea de recuperación.

Para la tarea t2i, se puede ver que F-HN logra los mejores resultados en Flickr30K, mientras que, en MS-COCO, RN obtiene el mejor rendimiento de recuperación seguido de F-HN. Cabe destacar que HN obtiene sistemáticamente el peor rendimiento de recuperación. Este resultado puede atribuirse al problema descrito previamente en el capítulo 3, esto es que, HN tiene problemas para manejar las similitudes intra-modales. En efecto, esto se comprueba monitoreando las similaridades efectivas que considera el modelo durante el entrenamiento. En la Fig 4.1 se presenta la similari-

[3] El mejor es subrayado, y el mejor del grupo RN, HN, F-HN y M-HN es destacado en negrita.

Cuadro 4.2: Resultados de los experimentos para recuperación de imagen a texto.

		Flickr30K						MS-COCO					
Metric		ZS	RN	HN	M-HN	F-HN	TERAN	ZS	RN	HN	M-HN	F-HN	TERAN
Recall	@1	0.20	40.5	**48.2**	41.8	47.3	75.8	0.02	20.34	**23.66**	18.14	18.82	55.6
	@5	0.70	67.5	**77.1**	72.3	75.5	93.2	0.14	45.62	**48.84**	41.32	42.42	83.9
	@10	0.90	80.6	84.7	82.7	**85.0**	96.7	0.26	59.5	**61.18**	54.3	55.9	91.6
nDCG@25	Rouge-L	0.261	0.566	**0.594**	0.553	0.585	0.687	0.244	0.516	**0.531**	0.492	0.497	0.643
	Spice	0.076	0.480	**0.509**	0.460	0.503	0.614	0.044	0.475	**0.483**	0.433	0.448	0.606
Novelty@25	Rouge-L	0.559	0.803	**0.830**	0.803	0.821	0.911	0.569	0.769	**0.781**	0.758	0.759	0.875
	Spice	0.154	0.665	**0.700**	0.657	0.694	0.811	0.094	0.676	**0.685**	0.648	0.658	0.805
	Self-information	7.309	7.292	**7.345**	7.291	7.249	7.285	9.094	9.224	**9.306**	8.676	8.569	9.667

Cuadro 4.3: Resultados de los experimentos para recuperación de texto a imagen.

		Flickr30K						MS-COCO					
Metric		ZS	RN	HN	M-HN	F-HN	TERAN	ZS	RN	HN	M-HN	F-HN	TERAN
Recall	@1	0.06	32.9	28.5	30.6	**39.4**	59.5	0.01	16.48	9.88	14.31	**19.48**	42.6
	@5	0.58	65.1	58.6	59.9	**69.0**	84.9	0.08	40.5	26.2	34.76	**42.68**	72.5
	@10	1.12	76.8	71.3	71.4	**79.2**	90.6	0.16	**54.71**	37.76	46.23	54.7	82.9
nDCG@25	Rouge-L	0.389	0.617	0.603	0.599	**0.629**	0.686	0.358	**0.618**	0.581	0.595	0.615	0.682
	Spice	0.128	0.498	0.482	0.471	**0.517**	0.564	0.065	**0.558**	0.503	0.521	0.551	0.610
Novelty@25	Rouge-L	0.616	0.816	0.802	0.804	**0.830**	0.880	0.621	0.801	0.770	0.786	**0.804**	0.868
	Spice	0.211	0.660	0.642	0.634	**0.675**	0.721	0.120	**0.720**	0.668	0.689	0.715	0.778
	Self-information	4.684	**5.187**	4.959	5.186	5.145	5.258	5.978	**7.357**	6.734	7.133	7.334	7.576

dad por épocas de entrenamiento, donde la línea continua corresponde a la similitud cross-modal, es decir, pares positivos o ground truth extraídos de los datos, mientras que, la línea discontinua corresponde a la similitud intra-modal entre pares no observados en los datos. Allí se comprueba que HN obtiene una $s(\bar{c}, c)$ por encima de $s(i, c)$ durante todo el proceso de aprendizaje. Por el contrario, tanto F-HN como M-HN requieren solo algunas épocas de entrenamiento para corregir este problema. Además, notar que F-HN es claramente más efectivo. Por otro lado, el modelo RN suele ser competitivo y robusto frente a valores atípicos (como descripciones que describen bien a más de una imagen) [6], porque de todas maneras serán muestreados con alta probabilidad negativos que son más similares o duros (en el sentido del hard negative) que el 90 % de todo el conjunto de entrenamiento [6]. Por supuesto, este efecto es proporcional al tamaño del mini-batch, y, por tanto, se verá reforzado por el utilizado en esta investigación.

4.3.2. Rendimiento de recuperación intra-modal.

Para tareas de recuperación intra-modal, F-HN logra el mejor rendimiento de recuperación según el nDCG (excepto para t2t en MS-COCO donde domina RN). Se debe tener en cuenta también que los resultados de rendimiento son siempre mayores para las tareas intra-modales en comparación a las tareas cross-modales. En particular, las métricas de recuperación intra-modal superan los resultados estado del arte en términos de nDCG [18] cross-modal. Más adelante se ejemplificará los beneficios de utilizar un modelo de recuperación cross-modal que incluye el manejo intra-modal por sobre uno que no lo hace. Gracias a un razonamiento y alineamiento

de información granular del contenido, TERAN logra resultados superiores en casi todos los casos. Sin embargo, F-HN reduce claramente la diferencia con TERAN, especialmente en las tareas intra-modales, e incluso para MS-COCO con la medida de relevancia Spice F-HN empata a TERAN, a pesar de utilizar representaciones globales o únicas del contenido. Por lo tanto, como trabajo futuro se propone aplicar las funciones de pérdida o loss propuestas con un modelo atencional, más detalles de este punto se expondrán en el capítulo 5.

Con respecto al enfoque ZS para la tarea i2i, si bien los márgenes se reducen en comparación a los resultados cross-modales, de todas formas el utilizar representaciones globales sin un ajuste de entrenamiento cross-modal resulta en un sistema menos preciso en todos los casos. En contraste, para la tarea t2t, el enfoque ZS resulta un baseline competitivo, incluso superando al estado del arte cross-modal para Flickr30K con la medida Rouge-L. Este resultado sugiere que el beneficio extraído de un aprendizaje de recuperación cross-modal es asimétrico en la modalidad, logrando mayores beneficios para la modalidad visual con el apoyo de la modalidad textual, en contraste del caso inverso. Este punto también será expuesto de nuevo en el capítulo 5.

Cuadro 4.4: Resultados de los experimentos para recuperación de imagen a imagen.

		Flickr30K						MS-COCO					
Metric		ZS	RN	HN	M-HN	F-HN	TERAN	ZS	RN	HN	M-HN	F-HN	TERAN
nDCG@25	Rouge-L	0.685	0.701	0.699	0.695	**0.703**	0.718	0.675	0.705	0.700	0.699	**0.708**	0.719
	Spice	0.545	0.574	0.574	0.563	**0.578**	0.582	0.576	0.618	0.609	0.609	**0.623**	**0.623**
Novelty@25	Rouge-L	0.918	0.925	0.924	0.923	**0.925**	0.930	0.910	0.931	0.930	0.930	**0.932**	0.936
	Spice	0.691	0.713	0.642	0.704	**0.716**	0.722	0.747	0.779	0.773	0.772	**0.783**	0.788
	Self-information	5.087	5.185	5.097	**5.219**	7.232	5.104	5.197	**7.393**	7.166	7.271	7.214	7.452

Cuadro 4.5: Resultados de los experimentos para recuperación de texto a texto.

		Flickr30K						MS-COCO					
Metric		ZS	RN	HN	M-HN	F-HN	TERAN	ZS	RN	HN	M-HN	F-HN	TERAN
nDCG@25	Rouge-L	0.764	0.716	0.726	0.646	**0.733**	0.758	0.696	**0.682**	0.671	0.586	0.662	0.714
	Spice	0.671	0.635	**0.645**	0.511	**0.645**	0.677	0.650	**0.650**	0.633	0.424	0.587	0.667
Novelty@25	Rouge-L	0.953	0.941	0.945	0.918	**0.946**	0.953	0.935	**0.933**	0.928	0.904	0.927	0.940
	Spice	0.789	0.766	0.642	0.652	**0.767**	0.795	0.837	**0.837**	0.826	0.688	0.805	0.848
	Self-information	7.471	7.455	7.342	7.510	**7.528**	7.292	9.568	9.613	9.401	9.771	**9.795**	9.670

4.3.3. Resultados de novedad

Para la tarea t2i, el modelo F-HN obtiene las mejores puntuaciones nDCG con sesgo de novedad, excepto para MS-COCO con Spice, donde ocupa el segundo lugar tras RN. Utilizando la medida de novedad basada en self-information e ignorando el modelo RN, F-HN y M-HN logran la mejor puntuación de novedad en MS-COCO y Flickr30K, respectivamente. Vale la pena señalar que HN obtiene el peor puntaje de novedad en todas las métricas en ambos conjuntos de datos (excepto con Spice en Flickr30K). Sin embargo, para la tarea i2t, HN supera a los otros modelos. Para esta tarea, F-HN logra los segundos mejores resultados con respecto a las métricas nDCG

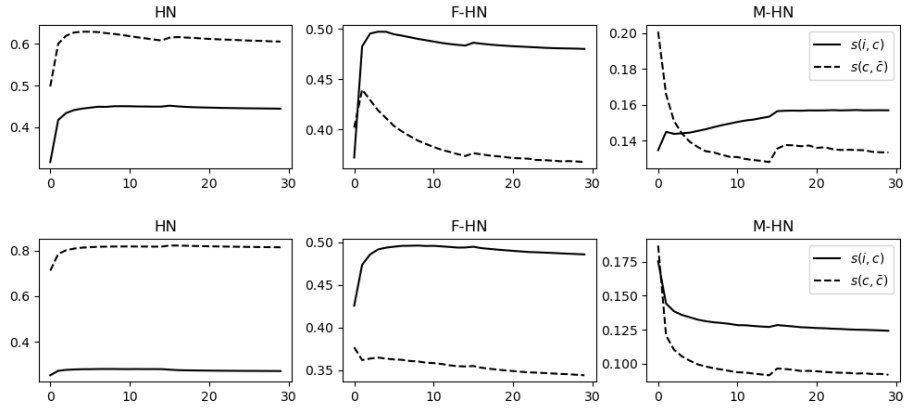


Figura 4.1: Similaridad promedio por época de entrenamiento en Flickr30K (arriba) y MS-COCO (abajo).

con sesgo de novedad Flickr30K. RN también produce resultados competitivos con respecto a las tres métricas de novedad introducidas en este trabajo, obteniendo a menudo mejor novedad que los modelos con mejores puntuaciones nDCG semántico. Este resultado se asemeja al clásico trade-off entre los principios de exploración y explotación (o novedad-precisión) [1].

Al considerar las tareas de recuperación intra-modal, F-HN obtiene casi siempre las puntuaciones de novedad más altas. En la tarea i2i, logra las mejores puntuaciones nDCG con sesgo de novedad en ambos conjuntos de datos. En la tarea t2t, F-HN supera a los otros modelos con respecto a la puntuación de novedad basada en self-information también en ambos conjuntos de datos, y logra las mejores puntuaciones de nDCG con sesgo de novedad en Flickr30K. Una vez más, vale la pena señalar que para ambas tareas de recuperación intra-modal, HN obtiene las peores puntuaciones de novedad en relación con la métrica de self-information.

Similarmente, a lo visto en términos de rendimiento, los resultados de novedad ilustran que los enfoques de aprendizaje cross-modal que incluyen componentes intra-modales pueden conducir a sistemas de recuperación de imágenes más novedosos.

4.3.4. Análisis cualitativo

Se considerará un ejemplo de recuperación cross e intra modal para ilustrar los beneficios de considerar F-HN en comparación a HN, particularmente cuando el objetivo es recuperar una imagen. En la Fig B.1 se presenta la imagen de referencia, la cual es descrita por los cinco captions presentes en el cuadro 4.6. A partir de esta tabla es posible notar que F-HN es superior a HN en todos los escenarios posibles.

Al considerar la imagen B.1 como consulta, se presenta las mejores cinco recuperaciones en las figuras B.3 y B.2 a partir de los modelos F-HN y HN, respectivamente. Se puede ver que los resultados son muy similares entre sí, con predominantemen-

te perros de color blancos y negros saltando una valla, tal como en la imagen de consulta. De hecho, la recuperación visual top 1; 2, denominaré esta como la imagen ∘; y 4 en la Fig B.3, son de las más similares a la consulta. En comparación al modelo HN, la imagen ∘ se pasa a top 3, y el top 4 anterior no es recuperado. Esto muestra empíricamente que la corrección del entrenamiento propuesto puede integrar recuperaciones relevantes. Sin embargo, el modelo F-HN no descarta totalmente la recuperación basada en información semántica que aporta la modalidad textual, ya que, por ejemplo, la imagen de referencia es descrita por el caption 4 (Q4) como una escena de un perro negro con blanco, lo cual se refleja en el top 3 de la Fig B.3, que fuera de ese punto, no comparte más contenido visual con la consulta, pues el fondo de la escena es totalmente distinto. Este argumento se ve potenciado al notar que esta imagen se recupera con frecuencia a partir de la modalidad textual, por tanto, se infiere que esta recuperación se apalanca de la semántica para ser recuperada.

Las recuperaciones a partir de texto con (F-HN) - (HN) son presentadas en las siguientes figuras B.5 - B.4, B.7 - B.6, B.9 - B.8, B.11 - B.10 y B.13 - B.12 con respecto a cada uno de los captions de la tabla. Para F-HN en cuatro de las cinco recuperaciones son un éxito y se recupera la imagen de referencia, además que usualmente se recuperan las imágenes relevantes mencionadas anteriormente. En contraste, HN nunca logra recuperar la imagen de referencia dentro de su top cinco, y de las imágenes más relevantes mencionadas anteriormente solo se recupera la imagen ∘. Esto se puede explicar debido a que la recuperación es susceptible a los términos usados en la descripción, lo cual genera variabilidad en función de ciertos conceptos claves presentes en la consulta, lo que lleva a la recuperación de imágenes no tan similares. Por ejemplo, se deduce que los sustantivos *jump poles*, *striped gate* y *barrier* son clave para recuperar la imagen ∘, en las figuras B.4 (top 5), B.8 (top 4) y B.12 (top 5), respectivamente. Mientras que para los otros dos captions faltantes, los sustantivos *hurdle* y *obstacle* aumentan la variabilidad de la recuperación en las figuras B.6 y B.10, respectivamente. Es posible deducir que la acción de salto, es un concepto semántico que tiene una influencia importante en las recuperaciones. Por ejemplo, se suelen recuperar perros saltando, pero con una pelota, lo que no se relaciona visualmente al objetivo. En otras palabras, se recuperan imágenes relacionadas semánticamente, pero a costa de perder precisión desde una perspectiva de similitud visual. En particular, hay una imagen que aparece en todas las recuperaciones cross-modales con HN, pero no en las intra-modales, esta es el top 1 en las figuras B.4, B.8, B.10 y B.12, o top 3 en la figura B.6. Dicha imagen es recuperada posiblemente porque guarda relaciones con la imagen de referencia a través de la modalidad textual, pues en la escena efectivamente hay un perro negro con blanco saltando, pero deduzco que, especialmente debido a que hay una valla detrás de él, la que es nombrada en las descripciones de dicha imagen, por ejemplo en *dog plays catch with a white ball near a wooden fence*, lo que podría inducir un acercamiento a nivel semántico con los sustantivos utilizados para describir lo que

Cuadro 4.6: Ejemplos de recuperaciones medidos por nDCG@25 con Rouge-L (R) y Spice (S).

Query	HN (R) - (S)	F-HN (R) - (S)
Q1: A dog jumps over an obstacle outside	0.500 - 0.549	0.687 - 0.675
Q2: A dog jumping over high jump poles at show	0.564 - 0.615	0.819 - 0.861
Q3: A small dog jumps over a striped gate	0.620 - 0.558	0.761 - 0.682
Q4: A black and white dog jumps over a hurdle	0.641 - 0.609	0.779 - 0.812
Q5: A small dog leaps a barrier	0.635 - 0.616	0.753 - 0.743
Q: Image B.1	0.781 - 0.775	0.804 - 0.777

está saltando el perro en la imagen de referencia (mencionados anteriormente), pero como a nivel visual no se parece, no tiende a aparecer en las recuperaciones ni intra ni cross modales con F-HN (excepto por el top 5 en la Fig B.13).

Capítulo 5

Conclusión y trabajo futuro

5.1. Conclusión

En esta investigación se exploraron los beneficios de la evaluación y entrenamiento cross-modal en tareas de recuperación intra-modal. Se dedujo que el entrenamiento cross-modal usando la triplet loss con hard negative mining puede llevar a relaciones de orden de similitud intra-modales inconsistentes en el espacio latente que es utilizado para la recuperación. Ello sirve como motivación para considerar nuevas alternativas al entrenamiento clásico. En efecto, para lidiar con este problema, se propusieron dos nuevas funciones de pérdida que consideran las relaciones de orden de similitud entre elementos cross e intra modales. La evaluación propuesta de estas técnicas fue más allá de la relevancia binaria, considerando tanto el contenido y la novedad de los elementos recuperados. Cabe recalcar que la metodología de evaluación no requiere de etiquetas ni para el entrenamiento ni para la evaluación. Además, los métodos propuestos son independientes de la manera en que se codifica o representa el contenido de la modalidad visual y textual, por lo que se pueden aplicar en diversas tareas o áreas.

Partiendo por la tarea de recuperación cross-modal, resultados experimentales indican que el método propuesto F-HN, que extiende la cross-modal triplet loss para imponer relaciones de orden de similitud coherente entre elementos cross e intra modales, produce mejoras en la recuperación de texto a imagen en comparación con el entrenamiento cross-modal convencional. Este método suele ocupar el segundo lugar en la recuperación de imágenes a texto, donde el enfoque convencional es una baseline más desafiante. El segundo método propuesto M-HN, que utiliza similitudes intra-modales para reemplazar el hiperparámetro de margen requerido por el enfoque convencional, produjo resultados más variados, pero la mayoría de las veces mejorando los resultados del método clásico. Sin embargo, es prometedor para aplicaciones con limitaciones de tiempo donde el ajuste de hiperparámetros es problemático. Con respecto a tareas de recuperación intra-modales, los experimentos en tareas de recuperación de imagen a imagen y de texto a texto revelaron que el método F-HN

es particularmente adecuado para tareas intra-modales, ya que a menudo proporciona listas de recuperación más precisas y novedosas. En general, estos hallazgos resaltan la importancia de considerar las similitudes intra-modales en el aprendizaje cross-modal, especialmente cuando la tarea implica recuperar elementos dentro de la misma modalidad, pero la relevancia debe determinarse utilizando otra modalidad como alternativa a no tener retroalimentación explícita.

5.2. Trabajo futuro

A partir de esta tesis se abren varias líneas de investigación que requieren tanto un estudio teórico como experimental. Las principales líneas a considerar son:

- Investigar el impacto de los métodos propuestos con encoders visuales y textuales más avanzados.
- Evaluar el efecto de incluir relaciones de similitud intra-modales para otras técnicas de muestreo negativo más allá del hard negative.
- Estudiar la relación entre la dimensionalidad del espacio latente y la probabilidad de inconsistencias intra-modales.
- Medir el rendimiento de la combinación de los métodos propuestos, por ejemplo, incorporando márgenes intra-modales (como en M-HN) en la formulación del método F-HN.
- Explorar el uso de hard negatives intra-modales para el entrenamiento cross-modal.
- Experimentar la calibración del margen por cada componente de la loss propuesta.
- Reparar la asimetría de los resultados dependiendo de la modalidad. Además de experimentar con datasets con diferentes proporciones de datos por modalidad.

Se expondrán más detalles para los dos puntos que se consideran más relevantes.

Particularmente, el primer punto podría conectarse con estudios sobre mecanismos atencionales para promover embeddings contextuales en escenarios multimodales. En detalle, TERAN utiliza una arquitectura neuronal tipo transformers para representar el contenido en cada modalidad, dicho método requiere del alineamiento entre los embeddings específicos entre regiones de interés de la imagen y palabras que componen el texto asociado. Así, el cálculo de similitud entre una imagen y un texto recae en un mecanismo atencional, específicamente, la similitud se calcula con la técnica *max-over-regions sum-over-words pooling* de la matriz de similitudes entre

regiones y palabras. Por lo tanto, la adaptación del método F-HN, implicaría reforzar el contexto cross-modal que provee la arquitectura de transformers, a partir del ajuste del contenido intra-modal. Por ejemplo, con respecto a la restricción visual en F-HN, ante una consulta de un texto, utilizando la técnica de pooling mencionada, se debería ajustar la similitud entre todos las regiones visuales más parecidas entre la imagen que corresponde a la consulta y el hard negative, inventando la técnica *max-over-regions sum-over-regions pooling*, con el objetivo de que dichas similitudes sean menor que la similitud entre las palabras que componen al texto de consulta y sus correspondientes regiones en la imagen. Esto tiene sentido, pues por el contexto, ambos representan el mismo concepto o significado en distintas modalidades, mientras que las regiones más parecidas del par negativo están en un contexto diferente. Se puede pensar en un caso de uso para clarificar este argumento, por ejemplo, ante una consulta que menciona la palabra pelota, y si la imagen hard negative también tiene una pelota parecida a la imagen que corresponde a la consulta, entonces el método F-HN impulsará un contexto apalancado de la información visual, para que las pelotas tengan una menor similitud en comparación a la palabra pelota, debido a que están en contextos diferentes y, por tanto, deberían ser representados de manera diferente, tal como la arquitectura transformer lo hace asignando representaciones distintas a palabras dependiendo de su contexto en tareas de NLP.

Si se extiende la hipótesis de este trabajo a modelos como TERAN, entonces la corrección propuesta promovería una mayor coherencia del contexto al momento de ajustar la información granular, ello debería permitir representaciones más discriminativas, lo que implicaría el diseño de sistemas de recuperación más precisos y novedosos.

Ahora, con respecto al último punto, los resultados experimentales de este trabajo muestran una asimetría en el beneficio que se logra al introducir correcciones intra-modales en el entrenamiento cross-modal, donde, la recuperación texto a imagen es más beneficiada en comparación a la recuperación de imagen a texto. Una hipótesis es que la multiplicidad de cinco textos por imagen en los datos es un factor relevante, ya que el éxito de la recuperación tiene cinco opciones, lo que podría implicar que pueda ser beneficioso para el espacio latente que haya representaciones de más textos compactos, como se evidencia empíricamente en la similitud promedio durante el entrenamiento (Fig 4.1) del método clásico HN en contraste de las propuestas. Manejar este fenómeno en los métodos propuestos es una línea abierta de investigación como consecuencia de los avances de esta tesis.

Bibliografía

[1] O. Berger-Tal, J. Nathan, E. Meron, and D. Saltz. The exploration-exploitation dilemma: A multidisciplinary framework. *PloS one*, 9:e95693, 04 2014.

[2] Y. Z. e. a. Chen Ma, Liheng Ma. Probabilistic metric learning with adaptive margin for top-k recommendation. *Proc. ACM SIGKDD 2020*, 2020.

[3] K. e. a. Clarke, Charles L.A. Novelty and diversity in information retrieval evaluation. SIGIR '08, page 659–666, New York, NY, USA, 2008. ACM.

[4] T.-T. Do, T. Tran, and e. a. Reid, Ian. A theoretically sound upper bound on the triplet loss for improving the efficiency of deep distance metric learning. In *IEEE CVPR*, pages 10404–10413, 2019.

[5] S. R. Dubey. A decade survey of content based image retrieval using deep learning. *IEEE Trans. on Circuits and Systems for Video Technology*, 32:2687–2704, 2020.

[6] J. R. K. Fartash Faghri, David J. Fleet and S. Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *Proc. BMVC*, 2017.

[7] W. Ge, W. Huang, D. Dong, and M. R. Scott. Deep metric learning with hierarchical triplet loss. In *ECCV 2018*, pages 272–288. Springer, 2018.

[8] F. L. Gewers, G. R. Ferreira, H. F. D. Arruda, F. N. Silva, C. H. Comin, D. R. Amancio, and L. D. F. Costa. Principal component analysis. *ACM Computing Surveys*, 54(4):1–34, may 2021.

[9] Y. Gong and G. Cosma. Improving visual-semantic embeddings by learning semantically-enhanced hard negatives for cross-modal information retrieval. *Pattern Recognition*, 137:109272, 2023.

[10] A. Gordo, J. Almazan, J. Revaud, and D. Larlus. End-to-end learning of deep visual representations for image retrieval. *Int. Journal of Computer Vision*, 124, 09 2017.

[11] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676, apr 2017.

[12] B. Kaya, Mahmut and H. Şakir. Deep metric learning: A survey. *Symmetry*, 11(9):1066, 2019.

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017.

[14] X. Li, J. Yang, and J. Ma. Recent developments of content-based image retrieval (cbir). *Neurocomputing*, 452:675–689, 2021.

[15] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[16] T.-Y. Lin, M. Maire, and e. a. Belongie, Serge. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[17] M. Mallea, R. Nanculef, and M. Araya. Enhancing intra-modal similarity in a cross-modal triplet loss. In A. Bifet, A. C. Lorena, R. P. Ribeiro, J. Gama, and P. H. Abreu, editors, *Discovery Science*, pages 249–264, Cham, 2023. Springer Nature Switzerland.

[18] N. Messina, G. Amato, and e. a. Esuli, Andrea. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(4):1–23, 2021.

[19] N. Messina, F. Falchi, A. Esuli, and G. Amato. Transformer reasoning network for image-text matching and retrieval. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5222–5229. IEEE, 2021.

[20] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, nov 1995.

[21] G. Molina, M. Mendoza, and e. a. Loayza, Ignacio. A new content-based image retrieval system for sars-cov-2 computer-aided diagnosis. In *MICAD 2021*, pages 316–324, 2022.

[22] M. J. Peter Anderson, Basura Fernando and S. Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398, 2016.

[23] R. Ren, S. Lv, and e. a. Qu, Yingqi. Pair: Leveraging passage-centric similarity relation for improving dense passage retrieval. pages 2173–2183, 01 2021.

[24] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

[25] L. Rosasco, E. D. Vito, A. Caponnetto, M. Piana, and A. Verri. Are loss functions all the same? *Neural Computation*, 16(5):1063–1076, 2004.

[26] E. Schubert. A triangle inequality for cosine similarity, 07 2021.

[27] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *IEEE CVPR*, pages 4004–4012, 2016.

[28] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu. Mpnet: Masked and permuted pre-training for language understanding. *NIPS*, 33:16857–16867, 2020.

[29] M. Tan and Q. V. Le. Efficientnetv2: Smaller models and faster training. *CoRR*, abs/2104.00298, 2021.

[30] Y. Tian, X. Yu, and e. a. Fan, Bin. Sosnet: Second order similarity regularization for local descriptor learning. pages 11008–11017, 06 2019.

[31] Y. T. Tony Ng, Vassileios Balntas and K. Mikolajczyk. Solar: Second-order loss and attention for image retrieval. *ArXiv*, 2020.

[32] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023.

[34] M. Wang and W. Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021.

[35] Z. Wang, Y. Wang, and e. a. Dong, Bo. Adaptive margin based deep adversarial metric learning. In *IEEE BigDataSecurity/HPSC/IDS 2020.*, pages 100–108, 2020.

[36] J. Z. Weihua Chen, Xiaotang Chen and K. Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. *IEEE CVPR*, pages 1320–1329, 2017.

[37] Y. Wu, S. Wang, and Q. Huang. Online asymmetric similarity learning for cross-modal retrieval. In *IEEE CVPR*, pages 3984–3993, 2017.

[38] Y. Wu, S. Wang, and Q. Huang. Online fast adaptive low-rank similarity learning for cross-modal retrieval. *IEEE Transactions on Multimedia*, 22(5):1310–1322, 2020.

[39] H. Xuan, A. Stylianou, X. Liu, and R. Pless. Hard negative examples are hard, but useful. In *ECCV*, pages 126–142, 11 2020.

[40] J. Yang, J. Duan, and e. a. Tran, Son. Vision-language pre-training with triple contrastive learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15650–15659, 2022.

[41] M. Ye, J. Shen, and e. a. Lin, Gaojie. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(6):2872–2893, 2021.

[42] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 12 2014.

[43] C. Zhao, X. Lv, and e. a. Zhang, Zhang. Deep fusion feature representation learning with hard mining center-triplet loss for person re-identification. *IEEE Transactions on Multimedia*, 22(12):3180–3195, 2020.

[44] T. Zhou, Z. Kuscsik, and e. a. Jianguo Liu. Solving the apparent diversity-accuracy dilemma of recommender systems. *PNAS*, 107:4511–4515, 2010.

Appendix A

Especificaciones de herramientas tecnologícas

A continuación se presentan las diversas herramientas y entornos para el desarrollo y prueba de software que se utilizaron durante este trabajo.

- Environment. Todos los modelos fueron implementados usando Python principalmente con la librería de tensorflow versión 2,10. El código completo que permite la reproducibilidad de los experimentos está disponible en `https://github.com/MariodotR/FullHN.git`.
- Hardware. Los experimentos fueron ejecutados en dos máquinas diferentes, en mi computador personal con una GPU 3060ti para Flickr-30K y em un serivdor con de la universidad que cuenta con una 1080ti para MS-COCO. En ningún caso, los tiempos de ejecución por modelo superan las 24 horas.

Appendix B

Ejemplos de recuperación

Figura B.1: Imagen de referencia o ground truth.

Figura B.2: Top 5 recuperaciones imagen a imagen con HN.

Figura B.3: Top 5 recuperaciones imagen a imagen con F-HN.

Figura B.4: Top 5 recuperaciones texto (Q1) a imagen con HN.

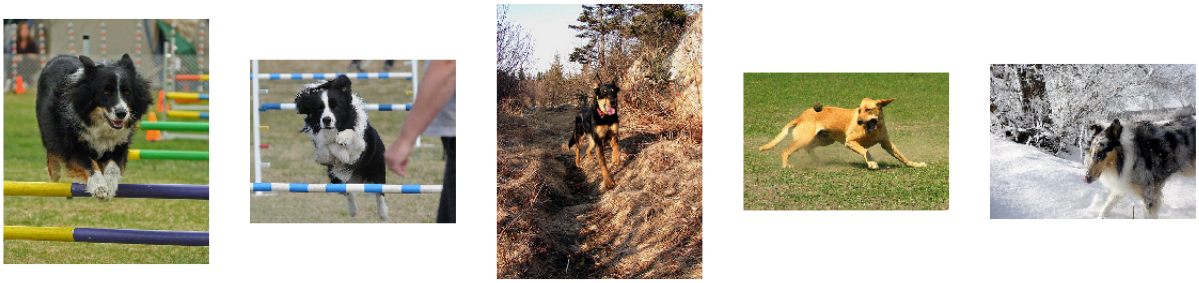


Figura B.5: Top 5 recuperaciones texto (Q1) a imagen con F-HN.

Figura B.6: Top 5 recuperaciones texto (Q2) a imagen con HN.

Figura B.7: Top 5 recuperaciones texto (Q2) a imagen con F-HN.

Figura B.8: Top 5 recuperaciones texto (Q3) a imagen con HN.

Figura B.9: Top 5 recuperaciones texto (Q3) a imagen con F-HN.

Figura B.10: Top 5 recuperaciones texto (Q4) a imagen con HN.

Figura B.11: Top 5 recuperaciones texto (Q4) a imagen con F-HN.

Figura B.12: Top 5 recuperaciones texto (Q5) a imagen con HN.

Figura B.13: Top 5 recuperaciones texto (Q5) a imagen con F-HN.