
This is the **published version** of the bachelor thesis:

Capdevila Estadella, Albert. *ViT-Explorer : Eina de visualització per a la interpretació de Vision Transformers*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Enginyeria Informàtica)

This version is available at <https://ddd.uab.cat/record/334441>

under the terms of the  license.

ViT-Explorer: Eina de visualització per a la interpretació de *Vision Transformers*

Albert Capdevila Estadella

29 de juny de 2026

Resum—Els *transformers* són una arquitectura que ha revolucionat tant la visió per computador com el processament del llenguatge natural, però la seva interpretabilitat s'ha estudiat més extensament en llenguatge que en visió. Per a reduir aquesta diferència, es proposa **ViT-Explorer**, una eina que combina anàlisi de dades i visualitzacions per a explorar les representacions internes d'un *Vision Transformer*. L'eina aporta tres contribucions de visualització: (a) un graf de l'espai latent reduït amb representacions d'atenció; (b) graelles de colors per visualitzar *embeddings*; i (c) mapes d'opacitat interactius per visualitzar l'atenció. Aquestes vistes es validen mitjançant tres casos d'ús, mostrant com evolucionen els *embeddings* des de trets de baix nivell fins a classes definides, com es separen *patches* idèntics en l'espai latent segons el context, i com els *heads* d'atenció presenten patrons diferents. Així doncs, ViT-Explorer fa una passa més en la interpretabilitat dels *Vision Transformers*.

Paraules clau—*Vision Transformers*, visió per computador, IA explicable, visualització d'*embeddings*, visualització d'autoatenció, visualització de l'espai latent, reducció de dimensionalitat, AlignedUMAP

Abstract—Transformers are a deep learning architecture that has revolutionized both Computer Vision and Natural Language Processing, however, their interpretability has been studied more extensively in language than in vision. To help bridge this gap, **ViT-Explorer** is proposed as a tool that combines visualization and data analysis techniques to explore Vision Transformers' internal representations across layers. It provides three main visualization contributions: (a) a low-dimensional latent space with graph-based attention representations; (b) a color-based embedding visualization; and (c) an interactive opacity map for attention display. These approaches are validated through three case studies, showing how embeddings evolve from low-level features to semantic classes, how identical patches are pulled away by their context in latent space, and how attention heads display different patterns across depth. Overall, ViT-Explorer is a further step toward explainable Vision Transformers.

Keywords—Vision Transformers, computer vision, explainable AI (XAI), embedding visualization, self-attention visualization, latent space visualization, dimensionality reduction, AlignedUMAP



1 INTRODUCCIÓ

EN els darrers anys, els *Transformers* [1] s'han consolidat com una de les arquitectures més rellevants en *deep learning*. Models de llenguatge com ChatGPT han contribuït a popularitzar aquesta tecnologia, però el seu funcionament intern continua sent complex i difícil d'entendre.

En visió per computador, els *Vision Transformers* (ViT) [2] estenen el mecanisme d'autoatenció al processament

d'imatges, assolint resultats competitius respecte a les arquitectures convolucionals.

Malgrat això, mentre que en el camp del processament del llenguatge natural s'han fet nombrosos esforços per interpretar el funcionament intern dels *transformers*, la comprensió dels *Vision Transformers* continua sent menys extensa. En particular, no s'ha visualitzat completament l'evolució de la representació de les parts de la imatge a través de les diferents capes, governada pels mecanismes d'atenció i per diversos tipus de transformacions.

Per aquest motiu, es planteja la hipòtesi que una eina que permeti navegar per les capes dels *Vision Transformers* i obtenir visualitzacions parcials de les seves representacions internes pot ajudar a reduir la incertesa sobre la interpretació

- E-mail de contacte: albertcapdev@gmail.com
- Menció: Computació
- Treball tutoritzat per: Maria Vanrell Martorell (Àrea de Ciències de la Computació i Intel·ligència Artificial)
- Curs 2025/26

d'aquests models i permetre dissenyar casos d'ús orientats a respondre preguntes sobre com emergeix el seu comportament.

Així doncs, els objectius d'aquest treball són:

Objectiu 1. Revisar les eines existents de visualització de *transformers* i estudiar els mètodes més rellevants per a l'anàlisi dels estats interns d'un *Vision Transformer*, amb l'objectiu d'identificar les tècniques més adequades per al desenvolupament de l'eina.

Objectiu 2. Construir una eina de visualització i anàlisi dels estats interns d'un *Vision Transformer* que permeti explorar l'evolució de les representacions des de l'entrada fins a la predicció final, representar el procés d'atenció sobre les diferents regions de la imatge i interactuar amb qualsevol model per observar el comportament dels seus estats interns.

Objectiu 3. Dissenyar i analitzar una sèrie de casos d'ús que permetin validar l'eina desenvolupada i, alhora, estudiar els mecanismes d'atenció i la representació de les diferents parts de la imatge al llarg de les capes del model.

L'eina construïda, anomenada ViT-Explorer, ofereix les següents utilitats:

- Un marc per a construir espais latents de baixa dimensionalitat (com ara PCA, t-SNE, UMAP, etc.) juntament amb una visualització en forma de graf dels pesos d'atenció.
- Una visualització, basada en color, de les representacions internes de les parts de la imatge.
- Un mapa d'opacitat que mostra interactivament l'atenció de cada part de la imatge respecte a una altra.

A més, s'han desenvolupat tres casos d'ús que han permès arribar a les següents observacions:

- L'estructura de l'espai latent evoluciona des de definir-se per trets posicionals i de baix nivell fins a representar classes completes.
- Parts idèntiques en imatges diferents evolucionen cap a posicions llunyanes en l'espai latent.
- Els *heads* d'atenció de les diferents capes mostren patrons diferenciats, alguns orientats a la posició espacial i d'altres especialitzats en estructures semàntiques com cares o ulls.

El treball s'ha organitzat de manera que el primer objectiu es tracta en els tres primers apartats, on s'introdueix el concepte de *Vision Transformer*, es fa una revisió bibliogràfica de les eines de visualització existents i s'explica la metodologia seguida tant pel que fa a la implementació de l'eina com als mètodes per analitzar els estats interns del model. Finalment, els dos darrers objectius es cobreixen a l'apartat de resultats, on es descriu ViT-Explorer i s'expliquen els casos d'ús realitzats.

2 CONTEXT: VISION TRANSFORMERS

En aquest treball, es fa referència a múltiples conceptes fonamentals dels *Vision Transformers*, motiu pel qual és necessari fer una breu definició i explicació del seu funcionament.

Un *Vision Transformer* (ViT) és una arquitectura d'aprenentatge profund que adapta el mecanisme d'atenció dels *transformers* per al processament d'imatges en tasques de visió per computador [3].

La unitat bàsica de representació en aquests models és l'*embedding*: un vector $e \in \mathbb{R}^{d \times 1}$ que codifica informació en un espai d -dimensional.

En un ViT, la imatge d'entrada es divideix en petits quadrats anomenats *patches* (o *tokens*), de dimensions $\mathbb{R}^{p \times p \times 3}$.

El primer pas del model és transformar cada *patch* en un *embedding* mitjançant una projecció lineal apresada durant l'entrenament. Així doncs, el conjunt inicial d'*embeddings* es pot representar com la matriu de N files:

$$\mathbf{E} = \begin{bmatrix} \mathbf{e}_{CLS}^\top \\ \mathbf{e}_1^\top \\ \vdots \\ \mathbf{e}_{N-1}^\top \end{bmatrix} \in \mathbb{R}^{N \times d}$$

on $e_1, \dots, e_{N-1}$ són els *embeddings* corresponents als *patches* de la imatge i e_{CLS} és l'*embedding CLS*, un vector entrenable que el model incorpora al conjunt per tal de capturar informació global de la imatge i posteriorment dur a terme la tasca de classificació.

A més, per incorporar informació sobre la posició dels *patches* en la imatge, a cada *embedding* se li suma un *embedding* de posició, de valors fixos o apresos.

Una vegada obtingut el conjunt inicial d'*embeddings*, aquests són transformats successivament al llarg de les capes del model segons l'esquema que es mostra a la Fig. 1.

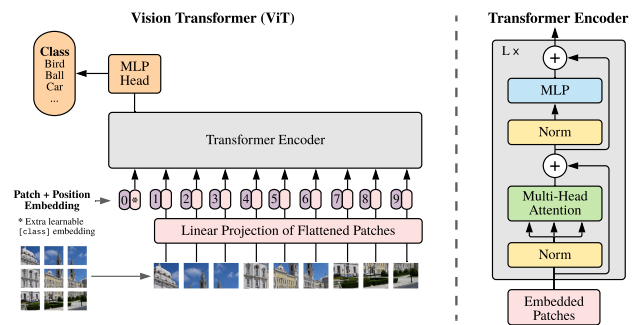


Fig. 1: Visió general d'un *Vision Transformer* (Figura reproduïda de [2]).

A continuació es descriuen els mòduls més complexos:

La **Multi-Head Attention** és el mecanisme que permet que els *embeddings* interaccionin i comparteixin informació entre si. Donat el conjunt d'*embeddings* d'entrada $E_{in} \in \mathbb{R}^{N \times d}$, es construeixen tres matrius mitjançant projeccions lineals:

$$\begin{aligned} Q &= E_{in} W_q^\top \in \mathbb{R}^{N \times m}, \\ K &= E_{in} W_k^\top \in \mathbb{R}^{N \times m}, \\ V &= E_{in} W_v^\top \in \mathbb{R}^{N \times v}. \end{aligned}$$

on $W_q \in \mathbb{R}^{m \times d}$, $W_k \in \mathbb{R}^{m \times d}$ i $W_v \in \mathbb{R}^{v \times d}$ són matrius de pesos apreses.

Les files $q_i^\top$, $k_i^\top$ i $v_i^\top$ de les matrius Q , K i V es poden interpretar conceptualment com a preguntes (*queries*), respostes (*keys*) i accions (*values*) corresponents a cada *embedding* e_i . Aleshores, l'atenció es calcula com:

$$E_{out} = \text{softmax}\left(\frac{QK^\top}{\sqrt{m}}\right) V = AV \quad (1)$$

Els elements de $QK^\top$ corresponen als productes escalars entre totes les parelles de *queries* i *keys*. Un valor elevat indica una **major alineació** entre vectors, i per tant una relació més forta entre els *embeddings* corresponents.

Després d'aplicar una normalització mitjançant *softmax*, s'obté una matriu de pesos, A , que representa l'atenció entre els diferents *embeddings* de la imatge (self-attention) i que determina la contribució de cada *value* (fila de V) a E_{out} . És a dir, si la *key* d'un *embedding* e_B s'alinea amb la *query* d'un altre *embedding* e_A , aleshores el vector de sortida corresponent a e_A tindrà més contribució del *value* de e_B que de la resta de *values*.

Aquest procés es repeteix H vegades en paral·lel utilitzant conjunts de matrius de pesos diferents, cadascuna de les quals defineix un *head* dins d'una mateixa capa.

Finalment, els vectors de sortida dels diferents *heads* es concatenen i es projecten de nou a la dimensió original dels *embeddings* mitjançant una matriu de pesos apresada (W^O). Els vectors resultants se sumen als *embeddings* originals mitjançant una connexió residual.

Tot seguit, als *embeddings* se'ls aplica un **Multi-Layer Perceptron (MLP)**: una projecció a una dimensionalitat superior, una transformació no lineal i una altra projecció a la dimensionalitat original.

Aquests dos processos es repeteixen al llarg de L capes, generant *embeddings* progressivament refinats (Vegeu Fig. 1 per a una estructura més detallada).

Finalment, la classificació es fa a partir de l'*embedding* CLS de l'última capa, que entra a un classificador lineal per obtenir la predicció.

3 ESTAT DE L'ART

A causa de la seva popularitat i a la necessitat d'entendre el seu funcionament, han aparegut nombrosos treballs dedicats a la visualització dels processos interns dels *transformers*. La majoria d'aquestes aproximacions s'han desenvolupat en l'àmbit del processament del llenguatge natural (NLP), mentre que la visió per computador ha rebut comparativament menys atenció.

En aquest apartat es revisen les visualitzacions més rellevants dels diferents components d'un *transformer*, amb exemples tant de llenguatge com de visió, ja que molts dels mecanismes interns són essencialment equivalents en ambdós camps.

3.1 Visualització d'*embeddings*

Per representar els *embeddings* a cada capa, alguns recursos redueixen la seva dimensionalitat mitjançant tècniques com **t-SNE** [4] o **UMAP** [5], i els mostren en un diagrama de dispersió on cada punt té assignat un color diferent segons alguna classificació determinada. Altres visualitzacions els

representen com a mapes lineals de valors, mostrant la intensitat de cada component de manera similar a un codi de barres [6].

A diferència d'aquests, **EL-VIT** [7] genera tantes imatges com dimensions tenen els *embeddings*, on cada imatge mostra el valor d'una posició concreta de tots vectors en una graella de colors reconstruïda segons els *patches* originals. Això permet visualitzar, simultàniament, l'evolució de cada una de les components dels *embeddings* al llarg de les capes.

Altres tècniques de comparació d'*embeddings* inclouen càlculs de distància en l'espai t-SNE [4] o graelles de *patches* on cada valor és la similitud cosinus entre un *embedding* seleccionat i el corresponent al *patch* [7].

3.2 Visualització de l'atenció

L'atenció és el mecanisme clau dels *transformers*, fet que explica la gran varietat de visualitzacions que existeixen dels seus components. Concretament, es poden dividir en dos nivells d'abstracció: els **caps (heads) d'atenció** i la **matriu de pesos d'atenció**.

3.2.1 Anàlisi dels heads d'atenció

Hi ha treballs que només visualitzen els *heads* d'atenció individualment [6], però l'àmplia majoria inclouen, en format de graella (capes $\times$ heads), una vista conjunta per tal de poder-los comparar [4, 8, 9].

Per identificar ràpidament quins *heads* són els més rellevants, hi ha propostes que calculen i mostren mètriques (per exemple, basades en la correlació dels pesos amb relacions lingüístiques conegudes [4, 5]), i altres que visualitzen els *heads* en un espai t-SNE per revelar patrons interessants [9, 10].

Dins del primer conjunt destaca **Dodrio** [5], ja que representa aquestes mètriques de manera intuïtiva utilitzant l'espai de color HCL (Vegeu Fig. 2).

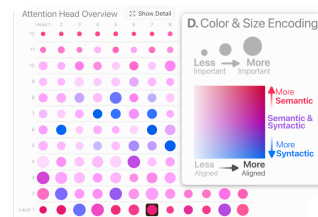


Fig. 2: *Heads* de color segons comportaments lingüístics mesurats amb mètriques (Figura reproduïda de [5]).

Dins del segon conjunt, **AttentionViz** [9] es diferencia per projectar, per cada *head*, els vectors *query* i *key* de múltiples imatges en els espais t-SNE, PCA i UMAP. Això genera patrons visuals en els quals la proximitat entre els punts reflecteix una intensitat d'atenció més alta.

3.2.2 Matriu de pesos d'atenció

La matriu d'atenció defineix com es distribueix l'atenció dins de cada *head* del model. Aquesta matriu es pot **visualitzar directament** amb un mapa de calor [6, 7], però també es pot interpretar com un conjunt d'interaccions entre *embeddings*. Així doncs, cada pes d'atenció entre dos *embeddings* es pot

representar com una aresta d'un **graf bipartit** [4, 8, 11], d'un graf radial [5] o d'altres variants (Vegeu Fig. 3).

En aquest cas, la visualització més rellevant és una de les proposades per **AttentionViz** [9], en la qual se **superposen fletxes** sobre els *patches* originals per mostrar les relacions d'atenció entre els seus *embeddings* corresponents. Concretament, per cada *embedding*, es dibuixa una fletxa cap a aquell altre *embedding* al qual ha dirigit més atenció. Si aquest és ell mateix o és l'*embedding* del token CLS, es mostra un cercle o un quadrat respectivament (Vegeu Fig. 3).

Aquesta visualització és especialment intuïtiva, ja que permet observar directament sobre la imatge com es distribueix l'atenció i quines regions influeixen més en altres.

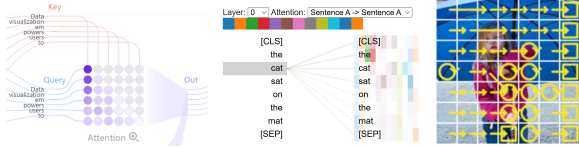


Fig. 3: Representacions de la matriu d'atenció: (A) Matriu de pesos. (B) Graf bipartit interactiu que mostra la intensitat dels pesos diferenciant els *heads* amb colors. (C) Fletxes d'atenció superposades. (Figures reproduïdes de [6, 8, 9]).

4 METODOLOGIA

El desenvolupament d'aquest treball ha requerit la selecció de les tecnologies utilitzades per implementar ViT-Explorer i la definició dels mètodes necessaris per a processar i transformar els estats interns d'un *Vision Transformer*. En aquesta secció es presenten les decisions preses en cadascun d'aquests àmbits.

4.1 Implementació de l'eina

En aquest apartat es justifica la selecció del model i les eines utilitzades per implementar la plataforma de visualització.

4.1.1 Selecció del model

Tot i que la plataforma desenvolupada s'ha dissenyat per ser aplicable a qualsevol model basat en l'arquitectura dels *Vision Transformers*, ha estat necessari seleccionar un model de referència sobre el qual implementar i validar les funcionalitats de visualització.

Amb aquest objectiu, s'han considerat diversos models basats en *Vision Transformers*. A la Taula 1 es mostra la comparació d'aquests segons diferents criteris rellevants per a la visualització i interpretació dels seus estats interns, com ara el nombre de *tokens* o la complexitat de l'arquitectura.

Finalment, s'ha seleccionat el model **ViT-B/32-AugReg-IN21k→IN1k**, ja que és un model de classificació amb una sortida de mil classes que divideix la imatge en *patches* de 32×32 píxels, generant un nombre reduït de *tokens*, i es troba implementat a la llibreria **timm**, àmpliament utilitzada en Visió per Computador.

El model està preentrenat en ImageNet, fet que permet centrar el treball en la visualització dels estats interns, evitant el cost computacional d'un entrenament des de zero.

¹Aquest model es limita a l'extracció de característiques i no incorpora cap classificació final; s'utilitza com a base per a altres models.

TAULA 1: COMPARACIÓ DE MODELS ViT

Tasca	Model	Criteris					
		Self-attention	Cross-attention	Nre. de tokens	Nre. de classes	Llibreria reconeguda	Interpretat
Extracció de trets ¹	ViT-B/32-IN21k	✓		50	-	✓	-
Classificació	ViT-B/16-IN1k	✓		197	1K	✓	Alta
	ViT-B/32-384-IN1k	✓		145	1K	✓	Alta
	ViT-B/32-IN21k → Food101	✓		50	101	✓	Baixa
	ViT-B/32-IN21k → Emotions	✓		50	8	✓	Baixa
	ViT-B/32-IN21k → CatsDogs	✓		50	2	✓	Baixa
	ViT-B/32-IN21k → Characters	✓		50	29	✓	Baixa
	ViT-B/32-IN1k (MLX backend)	✓		50	1K	✓	Alta
	ViT-B/32-AugReg-IN21k	✓		50	21K	✓	Alta
	ViT-B/32-AugReg-IN21k → IN1k	✓		50	1K	✓	Alta
Detecció d'objectes	YOLOS-S (ViT backbone)	✓		1829	80	✓	Alta
Detecció de posició	ViTPose-B (Simple)	✓		193	-	✓	Baixa
	ViTPose-B (Cross-Attention)	✓	✓	193	-	✓	Baixa
	ViTPose-S (Cross-Attention)	✓	✓	193	-	✓	Baixa
Segmentació	SAM-ViT-B	✓	✓	> 4096	-	✓	Alta

4.1.2 Desenvolupament de la plataforma

Amb l'objectiu d'implementar la plataforma de visualització ViT-Explorer, s'ha desenvolupat un conjunt d'eines en Python per a l'extracció dels estats interns del *Vision Transformer* i el seu processament segons la metodologia descrita en les seccions 4.2.1, 4.2.2 i 4.2.3 d'aquest treball.

Aquest conjunt de codi utilitza llibreries com **PyTorch**, **pandas** i **scikit-learn**, i s'encarrega de transformar les dades generades pel model en informació apta per a la visualització.

D'altra banda, la interfície de ViT-Explorer s'ha implementat utilitzant **Dash**, un *framework* que permet enllaçar fàcilment múltiples visualitzacions entre elles de manera interactiva i eficient mitjançant *callbacks*. Dash es basa en la llibreria **Plotly**, disponible tant en Python com en JavaScript. Aquest fet ha permès desenvolupar tota la lògica de visualització (i *callbacks* d'actualització) al costat del client, evitant la dependència de Python situat al servidor. Això ha millorat molt la velocitat d'actualització de totes les gràfiques, obtenint una interacció molt més fluida en la versió final.

Tot el codi desenvolupat, incloent-hi les eines d'extracció i la implementació de la interfície, es troba disponible al repositori del projecte: <https://github.com/albert-ce/vit-explorer>.

4.2 Mètodes de processament dels estats interns

Per a la visualització dels estats interns de ViT-Explorer, és necessari un processament previ que en permeti la representació i l'anàlisi. En aquest apartat es descriuen les transformacions utilitzades.

4.2.1 Reducció de dimensionalitat

Els *Vision Transformers* treballen amb vectors dins d'un espai d'alta dimensionalitat (concretament, el model seleccionat utilitza *embeddings* de 768 dimensions), els quals no són fàcilment interpretables per la ment humana. Per poder visualitzar-los i entendre'ls millor, és necessari reduir-ne la dimensionalitat.

Amb l'objectiu de construir una representació d'aquest espai, en aquest treball s'han capturat els *embeddings* de múltiples imatges al llarg de diverses capes i s'han projectat conjuntament mitjançant les següents tècniques:

- **PCA** Transformació lineal basada en els vectors propis de la matriu de covariàncies, maximitzant la variància projectada [12].
- **t-SNE**: Transformació no lineal que preserva les proximitats locals entre punts mitjançant distribucions de probabilitat [13].
- **UMAP**: Transformació no lineal que construeix un graf de veïnatge que preserva l'estructura local i parcialment la global [14, 15].

Entre aquestes tècniques, s'ha utilitzat també un mètode no explorat prèviament per a la reducció d'*embeddings* en aquest context:

- **AlignedUMAP**: Tècnica que consisteix a dividir les dades en particions i calcular reduccions UMAP independents, alineant-les entre elles mitjançant transformacions de Procrustes i mecanismes d'optimització [16].

En aquest treball, amb **AlignedUMAP**, els *embeddings* s'han agrupat segons la capa del model en la qual s'han capturat, construint així representacions independents per a cada nivell. Alinear aquestes permet fer comparables els *embeddings* al llarg del model, preservant la informació pròpia de cada capa però mantenint la idea d'un espai latent comú.

4.2.2 Extracció de les relacions d'atenció

L'atenció és el mecanisme central dels *transformers*, responsable de les interaccions entre *embeddings*. Aquestes relacions es poden representar amb una matriu $M \in \mathbb{R}^{N \times N}$ i es poden quantificar de diverses maneres. En aquest treball, s'han considerat les següents mètriques:

- **Pesos d'atenció (a_{ij})**: És la mètrica més intuïtiva i directa. Considera que la rellevància d'un *embedding* respecte a un altre és el valor a_{ij} corresponent de la matriu d'atenció $A = \text{softmax}\left(\frac{QK^T}{\sqrt{m}}\right)$ (Vegeu Eq. 1).
- **Normes de values ponderades (n_{ij})**: A diferència de l'anterior, aquesta mètrica [17] té en compte tots els elements que afecten l'impacte del mòdul d'atenció: els **pesos d'atenció**, els **values** i la **concatenació de heads**.

L'atenció és el producte de la matriu de pesos $A \in \mathbb{R}^{N \times N}$ i la matriu de values $V \in \mathbb{R}^{N \times v}$. En conseqüència, un *value* amb norma elevada i un pes baix poden contribuir de manera similar a un altre amb norma baixa però amb un pes d'atenció alt. A més, aquest procés es repeteix en cada *head*, pel que l'impacte a nivell de capa també depèn de la concatenació d'aquests.

Així doncs, per cada parell d'*embeddings* (i, j) en un *head* h , aquesta mètrica es calcula com:

$$n_{ij} = \|a_{ij} \mathbf{v}'_j\| \quad (2)$$

on $\mathbf{v}'_j$ és la fila j de la matriu

$$V' = VW_h^O \in \mathbb{R}^{N \times d}$$

on $W_h^O \in \mathbb{R}^{v \times d}$ és la part corresponent al *head* h de la matriu de concatenació de *heads* $W^O \in \mathbb{R}^{Hv \times d}$.

En definitiva, n_{ij} és la norma (com més alta, més rellevant) del vector procedent de l'*embedding* j que se sumerà a l'*embedding* i .

Per a la visualització d'aquesta mètrica, els valors s'han hagut de normalitzar entre 0 i 1 de manera similar a *Softmax*, de forma que la suma de totes les contribucions a un *embedding* concret sigui 1.

4.2.3 Tractament d'outliers en la visualització

En implementar els mètodes descrits anteriorment, s'ha observat que, a les capes finals del model, alguns *embeddings* amb poca rellevància semàntica sobresurten clarament respecte a la resta, condicionant fortament la representació global. Aquest fenomen es coneix com a "*artifact patches*" [18] i fa referència a *embeddings* redundants que actuen com a registres i que s'identifiquen per presentar normes vectorials atípiques a la sortida del model.

Com aquests artefactes dominen la representació dels *embeddings* i la de l'atenció, s'han considerat les següents opcions de filtratge:

- **Sense filtratge**: Es tracten tots els *embeddings* i tots els valors d'atenció de la imatge d'entrada.
- **Filtratge d'outliers superiors**: Es filtren els *embeddings* amb normes vectorials extremadament elevades, eliminant els artefactes de les representacions.
- **Filtratge de tots els outliers**: Es filtren tots els *embeddings* amb normes vectorials atípiques, fet que en molts casos elimina també l'*embedding* CLS.

Els efectes dels artefactes sobre l'estructura de l'espai latent es presenten amb més detall a la Secció 5.1.2, on es mostra que el filtratge d'aquests permet que la representació s'expandeixi i es defineixi millor (Vegeu Fig. 6).

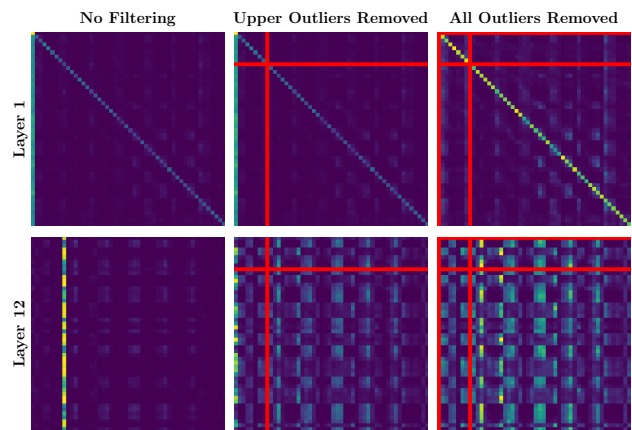


Fig. 4: Matrius de relacions d'atenció (Normes de values ponderades) per a dues capes del model sota les diferents condicions de filtratge. Les files i columnes filtrades es mostren en vermell.

Pel que fa a l'atenció, la Fig. 4 mostra les matrius d'atenció d'una mateixa imatge en dues capes del model (mètrica de normes de values ponderades). Sense filtratge, l'atenció es concentra en els *outliers*, generant patrons sorollosos. En

canvi, amb el filtratge complet i la renormalització de les representacions, els patrons d'atenció esdevenen clarament més interpretables i expressius.

5 RESULTATS

5.1 Eina desenvolupada: ViT-Explorer

ViT-Explorer és un *dashboard* interactiu per visualitzar els *embeddings* i els mecanismes d'atenció d'un conjunt d'imatges en les diverses capes d'un *Vision Transformer*. L'eina permet explorar aquesta informació de manera dinàmica mitjançant llindars, filtres i diferents opcions de selecció que faciliten l'anàlisi per part de l'usuari.

La informació es presenta en múltiples vistes coordinades, de manera que qualsevol interacció en una de les gràfiques es reflecteix automàticament en la resta. Això permet una exploració més clara, intuïtiva i completa dels resultats.

És una eina flexible i independent del model, adaptable a diferents *Vision Transformers* i conjunts d'imatges sense necessitat de modificar-ne l'estructura.

La interfície s'ha organitzat com un panell dividit en 5 seccions de vistes i selectors, que es poden veure a la Fig. 5 i que es descriuen a les següents seccions:

5.1.1 ① Controls i selectors

Aquesta secció agrupa els controls lliscants i selectors desplegable que permeten escollir les dades a representar i configurar la seva visualització:

- **Layer:** Selecciona en quina transició entre capes es volen mostrar els *embeddings*. Modificant el seu valor es pot veure l'evolució de l'espai d'*embeddings*.
- **Dimensions:** Selecciona el nombre de dimensions de la reducció. Les opcions són **2D** i **3D**.
- **Reduction:** Selecciona el mètode de reducció de dimensionalitat dels *embeddings*. Les opcions són **AlignedUMAP**, **t-SNE**, **PCA** i **UMAP** (Vegeu la Secció 4.2.1).
- **Attention Threshold:** Llindar doble per seleccionar la relació d'atenció màxima i mínima a mostrar.
- **Attention:** Selecciona la mètrica d'atenció. Les opcions són **Softmax** i **Weighted Value Norms** (Vegeu la Secció 4.2.2).
- **Images:** Selecciona les múltiples imatges a visualitzar del conjunt de dades.
- **Artifacts:** Selecciona l'opció de filtratge d'artefactes. Les opcions són **No Filtering**, **Upper Outliers Removed** i **All Outliers Removed** (Vegeu la Secció 4.2.3).

5.1.2 ② Graf de l'espai latent

En aquesta vista es representen els *embeddings* de les imatges seleccionades com a nodes d'un graf en un espai reduït 2D o 3D. A més, les relacions d'atenció entre *embeddings* es representen com arestes, on l'opacitat d'aquestes representa la seva intensitat.

En el cas de la representació en 2D, els *patches* originals es superposen als nodes, fet que permet associar ràpidament

el significat semàntic de l'*embedding* amb la seva posició dins l'espai latent (Vegeu Fig. 5 ②).

Els nodes són interactius: en seleccionar un *embedding*, la visualització s'actualitza per mostrar únicament les connexions entre aquest (*query*) i la resta d'*embeddings* de la imatge (*keys*). Quan un *embedding* s'atén a si mateix, es destaca amb una vora el gruix de la qual indica la magnitud d'aquesta atenció. Paral·lelament, la mida dels nodes varia segons l'atenció rebuda, facilitant la identificació dels elements més rellevants dins de cada *head*.

Aquesta visualització és especialment sensible a la presència d'*artifact patches*, per aquest motiu, és recomanable utilitzar les opcions de filtratge per visualitzar patrons més ocults (Vegeu Fig. 6).

5.1.3 ③ Vista i selecció de heads

En aquesta secció del *dashboard* es mostren, de manera reduïda, els grafs de l'espai latent de tots els *heads* del model. Això permet no només seleccionar el *head* que es vol analitzar dins la secció ②, sinó també visualitzar simultàniament els patrons que formen les arestes en cada *head*, facilitant la comparació i la identificació ràpida dels *heads* més rellevants o amb comportaments diferenciats.

5.1.4 ④ Graells d'*embeddings* de color

En aquesta vista es representa cada *embedding* mitjançant un color. Concretament, s'utilitzen les coordenades dels *embeddings* reduïts com les components *a* i *b* de l'espai de color *OKLAB* en el cas de 2 dimensions, o com les components *RGB* en el cas de 3 dimensions.

Així doncs, per a cada imatge es construeix la graella de *patches* original on cada cel·la mostra el color del seu *embedding* associat. A més, s'afegeix a la dreta de la graella una cel·la addicional corresponent a l'*embedding* del token CLS.

L'objectiu d'aquesta vista és representar de forma compacta i simple l'evolució semàntica dels *patches* i comparar la seva similitud de manera intuïtiva a través dels colors observats (Vegeu Fig. 7).

5.1.5 ⑤ Mapes d'opacitats

Aquesta vista mostra els *patches* de la imatge amb opacitats determinades per la mètrica d'atenció analitzada. És a dir, en passar el cursor per sobre d'un *patch*, es selecciona aquest i es visualitza, mitjançant variacions d'opacitat, el grau d'influència que tenen les altres parts de la imatge sobre la regió seleccionada. Si no hi ha cap selecció activa, es representa l'atenció mitjana rebuda per cada *patch*.

Com que les mètriques d'atenció presenten un gran rang dinàmic i estan fortament concentrades en valors baixos, s'aplica una transformació logarítmica per obtenir les opacitats. Aquesta transformació permet comprimir els valors alts i destacar millor les diferències entre valors baixos, millorant perceptualment la visualització.

És important destacar que les vistes ② i ⑤ estan sincronitzades: quan es fa una selecció en els mapes d'opacitats, es selecciona també automàticament l'*embedding* corresponent dins del graf de l'espai latent i a l'inrevés. Aquest fet permet que l'usuari relacioni de manera intuïtiva les dues representacions de la informació (Vegeu Fig. 8).

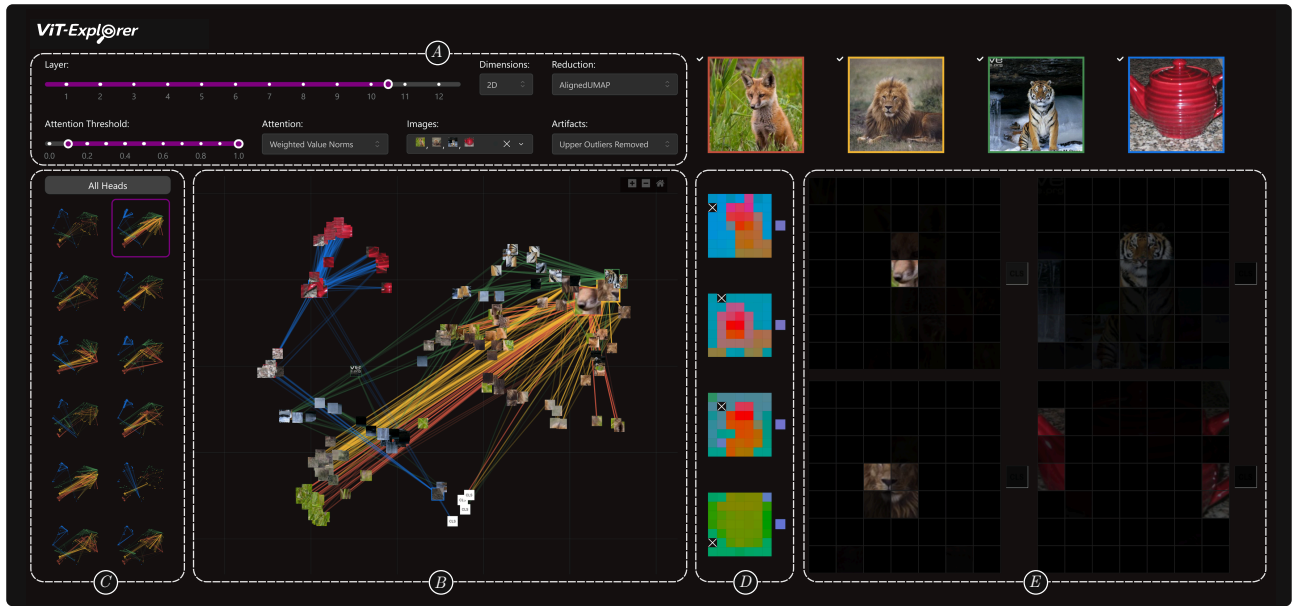


Fig. 5: ViT-Explorer: *Dashboard* interactiu per a la visualització dels estats interns d'un *Vision Transformer*



Fig. 6: Representació dels grafs de l'espai latent de múltiples imatges sota les diferents configuracions de filtratge (AlignedUMAP 2D).

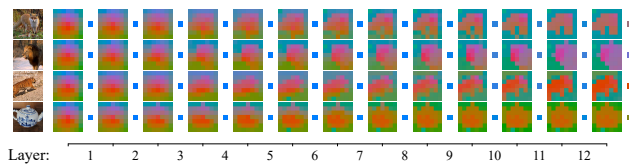


Fig. 7: Evolució de graelles de color al llarg de les etapes del model (AlignedUMAP 2D).

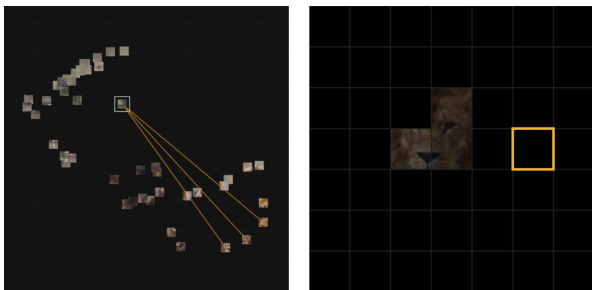


Fig. 8: Selecció sincronitzada d'un *embedding* al graf de l'espai latent (B) i al mapa d'opacitats (E).

5.2 Experiments i casos d'ús

Amb l'objectiu d'il·lustrar el funcionament de ViT-Explorer i recollir les observacions rellevants obtingudes a partir de la seva utilització, aquesta secció presenta tres casos d'ús.

En els tres casos s'ha treballat amb un subconjunt de 240 imatges d'ImageNet, format per quatre classes diferents (*fox*, *lion*, *tiger* i *teapot*), amb 60 imatges per classe. La selecció de les classes s'ha fet amb l'objectiu d'observar *embeddings* de *patches* similars, incloent-hi animals com el tigre i el lleó, que comparteixen formes globals però difereixen en color i textura, i contrastar-los amb elements completament diferents, com la tetera.

L'anàlisi de les representacions s'ha dut a terme en l'espai 2D resultant d'aplicar AlignedUMAP sobre els *embeddings* d'aquestes imatges en les diferents capes del model. S'han filtrat tots els *outliers* en els dos primers casos d'ús.

Els casos d'ús considerats analitzen: l'evolució semàntica de l'espai latent a partir de clústers d'*embeddings*, l'efecte del context en *patches* visualment idèntics en imatges diferents, i l'especialització en determinades característiques dels *heads* d'atenció.

5.2.1 Evolució semàntica a l'espai latent

El graf de l'espai latent de ViT-Explorer mostra les projeccions dels *embeddings* a les diferents capes del model. Aquesta representació varia progressivament al llarg del processament, amb canvis en la distribució i agrupació dels punts.

En aquest cas d'ús s'analitza l'estructura d'aquest espai a partir d'agrupacions dels *embeddings* reduïts, amb l'objectiu d'observar com s'organitza la representació semàntica al llarg de les capes del *Vision Transformer*.

Per a aquesta anàlisi, s'ha etiquetat una imatge representativa de cada classe a múltiples nivells d'abstracció (vegeu la Fig. 9).

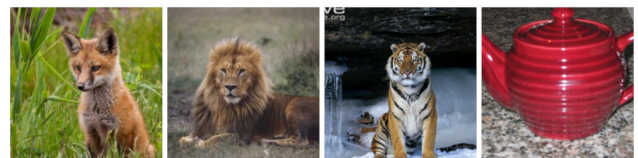


Fig. 9: Imatges representatives utilitzades per a l'etiquetatge

L'etiquetatge es basa en les següents categories semàntiques: **position** (*patch 1, patch 2...*), **low level** (*striped, red...*), **parts** (*eye, nose...*), **object** (*animal...*), **class** i **other** (*foreground, background...*).

A continuació, els *embeddings* de cada capa s'han agrupat en 49 clústers mitjançant K-Means. S'ha escollit $k = 49$ ja que cada imatge es divideix en 49 *patches*.

Seguidament, s'ha definit la puntuació $S_L^C(k)$ per analitzar les etiquetes predominants en aquests clústers. Aquesta mesura calcula, per cada categoria, la similitud cosinus entre totes les parelles d'*embeddings* d'un clúster. Concretament, la similitud es calcula sobre els vectors *one-hot* associats a les etiquetes de cada *embedding*. Per a la capa L_t , categoria C_f i clúster k_m , es defineix com:

$$S_{L_t}^{C_f}(k_m) = \frac{2}{n \cdot (n-1)} \sum_{i=1}^n \sum_{j=i+1}^n \frac{\vec{E}(\vec{x}_i) \cdot \vec{E}(\vec{x}_j)}{\|\vec{E}(\vec{x}_i)\| \|\vec{E}(\vec{x}_j)\|}$$

Finalment, per a cada categoria s'ha calculat la mitjana de les puntuacions dels clústers de cada capa, obtenint així l'evolució de les puntuacions al llarg del model (Vegeu Fig. 10).

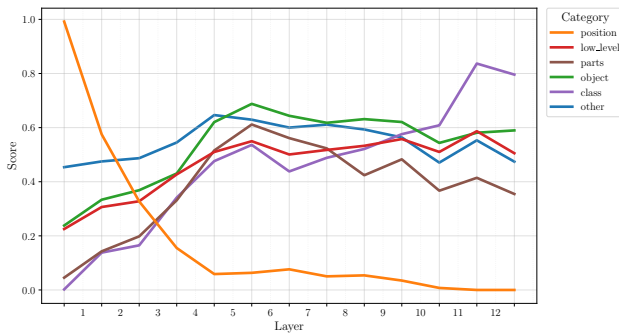


Fig. 10: Puntuacions d'estructura semàntica al llarg de les capes del model.

Els resultats mostren que l'estructura de l'espai projectat està inicialment dominada per les etiquetes de **posició** dels *patches*, les quals van perdent rellevància de manera pronunciada en les quatre primeres capes del model.

La rellevància de les característiques de **baix nivell** i de les etiquetes **d'objecte** va augmentant progressivament en les primeres capes i es manté estable a partir de la capa 5. En canvi, la puntuació de la categoria de **parts** augmenta inicialment però posteriorment disminueix, alhora que la categoria de **classe** segueix prenent rellevància. De fet, la definició de les classes de les imatges, que és la tasca del model, incrementa fins a dominar l'estructura final de l'espai, amb un augment pronunciat a la capa 11.

Aquesta evolució, des de la influència posicional fins a la definició de les classes, també es pot observar clarament en les graelles d'*embeddings* de color, que evolucionen des d'un gradient fins a la formació de siluetes dels elements (Vegeu Fig. 7).

En definitiva, tot i que aquests resultats són exploratoris i s'haurien de validar amb experiments més robustos, confirmen la utilitat de la projecció de ViT-Explorer amb AlignedUMAP i mostren una evolució clara de l'estructura de l'espai latent del model al llarg de les capes, amb indicis

d'una progressió jeràrquica similar a l'observada en xarxes convolucionals.

5.2.2 Contextualització semàntica de *patches* idèntics

ViT-Explorer permet seleccionar un *embedding* i fer el seguiment de la seva evolució al llarg de les capes del model, fet que fa possible analitzar si la seva representació varia segons el context.

L'objectiu d'aquest cas d'ús és utilitzar les visualitzacions desenvolupades per examinar si dos *embeddings* inicialment idèntics, tot i pertànyer a classes diferents, evolucionen de manera similar o si la seva trajectòria es separa en funció del context.

Per observar aquest fenomen, s'han construït dos parells d'imatges modificades (Vegeu Fig. 11). Les dues primeres imatges comparteixen el mateix *patch* corresponent a un ull de l'animal, mentre que les imatges restants comparteixen el mateix *patch* corresponent a una zona d'herba.

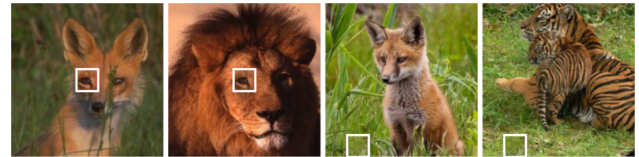


Fig. 11: Parells d'imatges alterades amb *patches* idèntics (destacats en blanc).

Aquestes imatges s'han inclòs al conjunt utilitzat per construir l'espai latent visualitzat a ViT-Explorer.

A la Fig. 12 es representen les trajectòries dels *embeddings* seleccionats al llarg de les capes del model, on l'evolució de cada *embedding* d'una capa a l'altra es mostra mitjançant vectors de diferents opacitats.

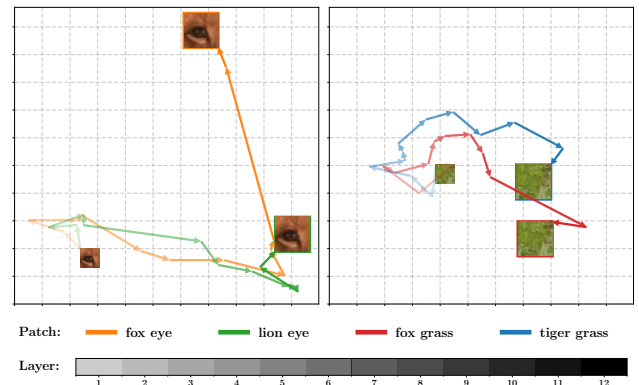


Fig. 12: Trajectòries dels *embeddings* seleccionats en l'espai 2D obtingut amb AlignedUMAP.

S'observa que, partint de la mateixa posició inicial, les trajectòries dels parells d'*embeddings* es mantenen properes fins a les capes finals, on es diferencien de manera més marcada. Aquest desplaçament és especialment pronunciat en el cas de l'ull, ja que la cara de l'animal és un element rellevant per a la classificació. En canvi, es mostra una separació menys pronunciada en el cas del *patch* d'herba, que té una menor càrrega semàntica en relació amb la classificació.

Concretament, s'observa que l'augment pronunciat de la distància entre els *embeddings* corresponents als ulls es produeix a la capa 11 del model. Per estudiar si aquesta

separació reflecteix un canvi semàntic en la representació, s'ha aplicat el cap de classificació del model sobre els dos *embeddings* abans i després d'aquesta capa. Tal com es mostra a la Taula 2, abans de la capa 11, la classe correcta no apareix entre les cinc prediccions més probables en cap dels dos casos, mentre que després de la separació passa a ocupar la primera posició en ambdós *embeddings*.

TAULA 2: PREDICCIONS TOP-5 PER ALS EMBEDDINGS D'ULLS D'ANIMALS

	Fox Eye		Lion Eye	
	Class	Prob.	Class	Prob.
Before L11	white wolf	0.478	Leonberg	0.132
	dhole	0.083	hyena	0.065
	coyote	0.071	American egret	0.035
	kit fox	0.042	African hunting dog	0.033
	Australian terrier	0.033	ibex	0.026
After L11	Class	Prob.	Class	Prob.
	red fox	0.788	lion	1.000
	kit fox	0.212	cougar	4.14e-13
	grey fox	4.54e-8	jaguar	1.33e-13
	Arctic fox	4.40e-8	leopard	2.98e-14
	coyote	1.35e-8	cheetah	4.36e-15

Així doncs, els resultats obtinguts en aquest cas d'ús indiquen que l'espai reduït de ViT-Explorer no només permet observar una separació entre els *embeddings* inicialment idèntics, sinó que també mostra aquesta separació quan realment es produeix un canvi semàntic en l'espai d'alta dimensionalitat.

5.2.3 Especialització dels *heads* d'atenció

Un altre dels aspectes que permet visualitzar ViT-Explorer són les relacions d'atenció produïdes en cada *head* d'atenció del model. Examinant les vistes de l'eina, s'ha observat que alguns *heads* presenten comportaments específics, com ara patrons posicionals, on l'atenció es concentra entre regions properes de la imatge. Alguns dels *heads* posicionals identificats són el H4L2, el H3L3, el H11L3 i el H1L4, on cada *embedding* atén principalment el següent de la seva fila o columna, el H12L2, que dirigeix l'atenció als *embeddings* veïns, i el H8L3, que manté l'atenció dins d'una mateixa columna.

Els *heads* posicionals són fàcils d'identificar, ja que es mostren com un graf de camins dirigits en la vista (B) i s'interpreten de forma directa als mapes d'opacitats (Vegeu Fig. 13).

En capes superiors, els mapes d'opacitats també suggereixen l'existència de *heads* que atenen parts específiques de les classes d'imatge, com ara cares d'animals (Vegeu Fig. 14), en contrast amb els patrons principalment posicionals observats en capes anteriors.

En conjunt, aquestes observacions indiquen que ViT-Explorer és útil per identificar i explorar l'especialització dels *heads*, des de patrons posicionals en capes inicials fins a patrons més semàntics en capes superiors.

També és rellevant assenyalar que, prèviament al desenvolupament de ViT-Explorer i amb l'objectiu d'analitzar l'especialització dels *heads*, s'han calculat els *patches* associats als *embeddings* més atesos en cada *head* segons diferents mètriques. Els resultats obtinguts són exploratoris,

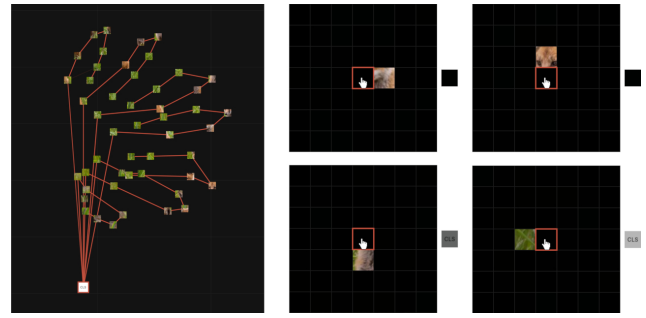


Fig. 13: Visualitzacions de *heads* posicionals: (A) Graf de camins dirigits (B) Mapes d'opacitats dels *heads* H4L2, H3L3, H11L3 i H1L4.



Fig. 14: Visualització de les atencions mitjanes rebudes al H11L10 (AlignedUMAP 2D, Weighted Value Norms, All Outliers Removed).

però indiquen un potencial clar per a possibles investigacions futures, ja que apareixen *patches* de colors uniformes i textures simples en els *heads* de les primeres capes i elements més complexos com rostres humans o ulls d'animals en *heads* de capes superiors (Vegeu la Secció A.1).

6 CONCLUSIONS

La motivació d'aquest treball ha estat contribuir a la interpretabilitat dels *Vision Transformers* mitjançant el desenvolupament de ViT-Explorer, una eina interactiva per explorar les seves representacions internes. Amb aquest propòsit, s'han desenvolupat tres visualitzacions interactives i s'han definit tres casos d'ús que han permès validar-ne la utilitat i estudiar diferents aspectes del funcionament intern d'aquests models.

En conjunt, aquest treball mostra com una eina de visualització pot facilitar la comprensió dels *Vision Transformers*, fet que obre la porta a noves línies de recerca centrades en la interpretabilitat d'aquests models, així com al desenvolupament de mètodes més avançats per analitzar i explicar el comportament de les seves representacions internes.

A nivell personal, més enllà de les contribucions tècniques del treball, aquest projecte ha suposat una oportunitat per aprofundir en la comprensió de l'arquitectura i del funcionament dels *Vision Transformers*, obtenint una intuïció més propera sobre el comportament dels seus estats interns. Igualment, ha permès adquirir experiència en el desenvolupament de visualitzacions web interactives amb llibreries com Dash i en el processament de dades necessari per construir representacions útils per a la interpretació de models d'IA.

6.1 Treball futur

Tot i que els objectius plantejats s'han assolit satisfactòriament, durant la recerca i el desenvolupament del projecte han aparegut diversos aspectes que podrien incorporar-se a l'eina o explorar-se en estudis futurs. Entre les possibles línies de treball futures destaca la integració de mètriques d'atenció existents, com l'*Attention Rollout* [19], o el desenvolupament de noves mètriques, com l'aplicació del principi de *Rollout* a les *Weighted Value Norms* en lloc d'utilitzar els pesos d'atenció. Així mateix, es podrien incorporar a l'eina mètodes per identificar les regions de la imatge que contribueixen de manera més directa a la decisió final del model [20].

També seria rellevant aprofundir de manera extensa i metòdica en els experiments de l'evolució semàntica a l'espai latent i el de l'especialització dels *heads* d'atenció, ja que en aquest treball s'han explorat de manera preliminar però mostren un clar potencial per a una anàlisi més profunda en treballs futurs.

ÚS DE LA IA GENERATIVA

En aquest treball, la intel·ligència artificial s'ha utilitzat per millorar la redacció i la correcció del text, per a la traducció de termes tècnics, i com a assistència puntual en la cerca preliminar d'articles i la resolució d'errors de programació, sempre sense afectar els continguts avaluable ni el desenvolupament conceptual del treball.

AGRAÏMENTS

Voldria agrair a la meua tutora, Maria Vanrell Martorell, el seguiment continuat i el seu suport durant l'elaboració d'aquest treball. La seva dedicació i les orientacions proporcionades han estat fonamentals per al desenvolupament i la finalització d'aquest TFG.

REFERÈNCIES

- [1] A. Vaswani et al., "Attention is All you Need," a *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. [adr. \(1\)](#).
- [2] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," a *International Conference on Learning Representations*, 2021. [adr. \(1, 2\)](#).
- [3] A. Torralba et al., *Foundations of Computer Vision* (Adaptive Computation and Machine Learning series). MIT Press, 2024, ISBN: 9780262378666. [adr. \(2\)](#).
- [4] V. Lal et al., "InterpreT: An Interactive Visualization Tool for Interpreting Transformers," a *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Online: ACL, 2021. [DOI \(3, 4\)](#).
- [5] Z. J. Wang et al., "Dodrio: Exploring Transformer Models with Interactive Visualization," a *Proceedings of the 59th Annual Meeting of the ACL and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, Online: ACL, 2021. [DOI \(3, 4\)](#).
- [6] A. Cho et al., "TRANSFORMER EXPLAINER: Interactive Learning of Text-Generative Models," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. [DOI \(3, 4\)](#).
- [7] H. Zhou et al., "EL-VIT: Probing Vision Transformer with Interactive Visualization," a *IEEE ICDMW 2023*, ISSN: 2375-9259, 2023. [DOI](#). cons. 24 de febr. de 2026 [\(3\)](#).
- [8] J. Vig, "A Multiscale Visualization of Attention in the Transformer Model," a *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Florence, Italy: ACL, 2019. [DOI \(3, 4\)](#).
- [9] C. Yeh et al., "AttentionViz: A Global View of Transformer Attention," *IEEE Transactions on Visualization and Computer Graphics*, 2024. [DOI \(3, 4\)](#).
- [10] Y. Li et al., "How Does Attention Work in Vision Transformers? A Visual Analytics Attempt," *IEEE Transactions on Visualization and Computer Graphics*, 2023. [DOI \(3\)](#).
- [11] B. Hoover et al., "exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models," a *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Online: ACL, 2020. [DOI \(4\)](#).
- [12] PCA, en. cons. 10 d'abr. de 2026. [adr. \(5\)](#).
- [13] TSNE, en. cons. 10 d'abr. de 2026. [adr. \(5\)](#).
- [14] UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction — *umap 0.5.8 documentation*. cons. 10 d'abr. de 2026. [adr. \(5\)](#).
- [15] A. Coenen et al., *Understanding UMAP*. cons. 10 d'abr. de 2026. [adr. \(5\)](#).
- [16] *How to use AlignedUMAP — umap 0.5.8 documentation*. cons. 22 de maig de 2026. [adr. \(5\)](#).
- [17] G. Kobayashi et al., "Attention is Not Only a Weight: Analyzing Transformers with Vector Norms," a *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: ACL, 2020. [DOI \(5\)](#).
- [18] T. Darcet et al., "Vision Transformers Need Registers," en, *International Conference on Learning Representations*, 2024. cons. 3 de juny de 2026. [adr. \(5\)](#).
- [19] S. Abnar et al., "Quantifying Attention Flow in Transformers," a *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: ACL, 2020. [DOI \(10\)](#).
- [20] H. Chefer et al., "Transformer Interpretability Beyond Attention Visualization," a *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. [DOI \(10\)](#).
- [21] G. Sanderson. "Attention in Transformers, Step-by-Step." Deep Learning, Chapter 6, cons. 3 de març de 2026. [adr.](#)
- [22] J. Alammari, *The Illustrated Transformer*, 2018. cons. 3 de març de 2026. [adr.](#)

APÈNDIX

A.1 Estudi exploratori de l'especialització de *heads* d'atenció

Amb l'objectiu d'analitzar diversos mètodes per definir les relacions d'atenció i explorar la possible especialització dels *heads* del model, s'han trobat els *patches* associats als *embeddings* més atesos del conjunt de dades.

Aquesta experimentació s'ha dut a terme de la següent manera. Donada la matriu de relacions d'atenció $M \in \mathbb{R}^{N \times N}$, on cada element representa el grau d'influència d'un *embedding* respecte d'un altre, s'ha calculat la suma per columnes per obtenir l'atenció total rebuda per cada *embedding*. És a dir, cada *embedding* té associat un valor d'atenció rebuda $a_i = \sum_{j=1}^N M_{ji}$.

Aleshores, per analitzar si el *head* d'atenció presenta una especialització, s'han extret els *patches* associats als 50 *embeddings* amb major atenció rebuda.

Aquest procediment s'ha aplicat sobre un subconjunt d'ImageNet de 5.000 imatges (5 imatges per classe) i per a les diferents mètriques d'atenció descrites a la Secció 4.2.2, entre d'altres. Les altres mètriques inclouen criteris com ara les normes dels *values* sense ponderar o els pesos d'atenció del model abans de l'aplicació del *softmax*.

Els resultats més rellevants es mostren a les Figs. 15 i 16.

S'observa que, en els *heads* d'atenció de la primera capa, els *patches* amb més atenció rebuda corresponen principalment a patrons de color uniforme o a vores verticals i horitzontals. Concretament, el *head* 1 destaca *patches* blancs i negres, el *head* 2, *patches* blaus i verds, i el *head* 4, *patches* associats a colors més diversos, com vermell, groc i blau.

En canvi, en una capa superior, com la capa 6, els *patches* més rellevants presenten textures més complexes o, fins i tot, en el cas del *head* 10, s'hi distingeixen clarament elements d'alt nivell semàntic, com cares humanes i ulls d'animals.

De fet, els *patches* resultants semblen indicar que cada *head* dona rellevància simultàniament a múltiples tipus de característiques. Per analitzar aquest fet, s'han agrupat els *patches* més rellevants aplicant l'algorisme K-Means sobre els seus *embeddings*. Els resultats més prometedors es mostren a les Figs. 17 i 18. Tot i això, la dificultat d'obtenir un nombre de clústers adequat que produís resultats consistents a tots els *heads* va ser el motiu pel qual no es va continuar amb aquesta línia d'anàlisi.

En conjunt, aquesta experimentació sobre l'especialització dels *heads* del model s'emmarca en una línia de recerca que queda fora de l'abast d'aquest treball de fi de grau. No obstant això, aquesta podria ser objecte d'estudi futur, ja que els resultats preliminars obtinguts mostren potencial.

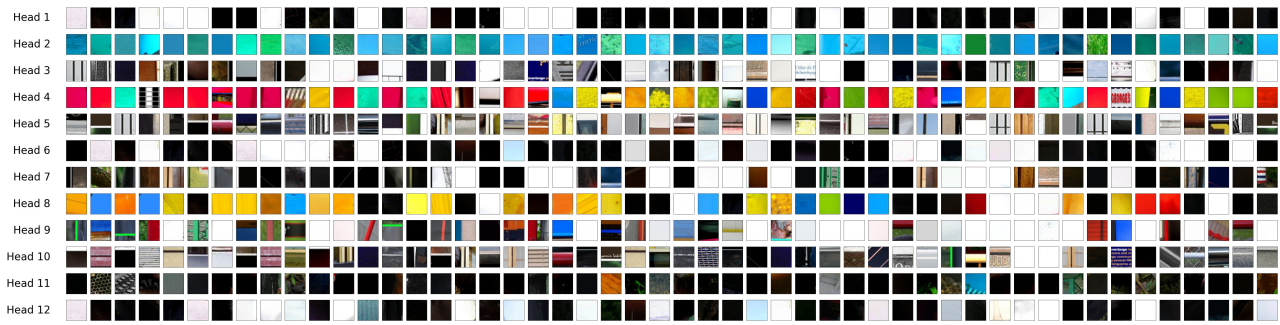


Fig. 15: Els 50 *patches* més atesos als *heads* de la capa 1 amb el criteri *Weighted Value Norms*.

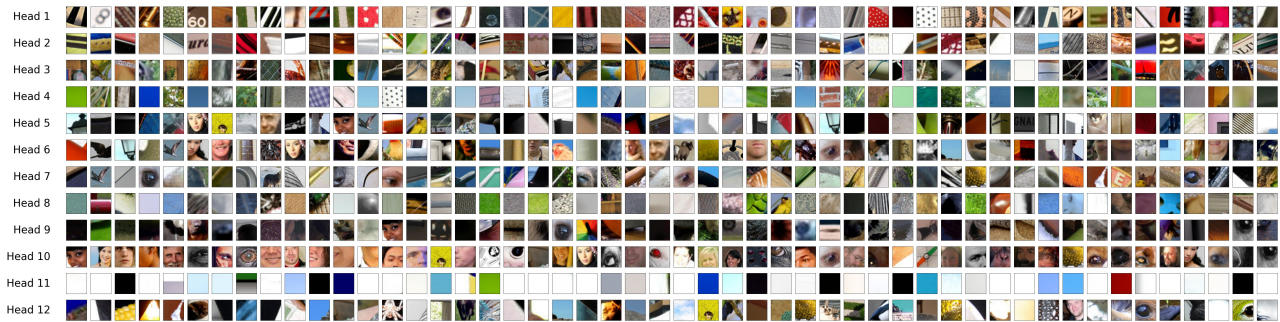


Fig. 16: Els 50 *patches* més atesos als *heads* de la capa 6 amb el criteri *Weighted Value Norms*.



Fig. 17: Agrupació amb K-Means dels 200 *patches* més atesos al H4L1 utilitzant el criteri *Softmax*. Cada fila correspon a un clúster.

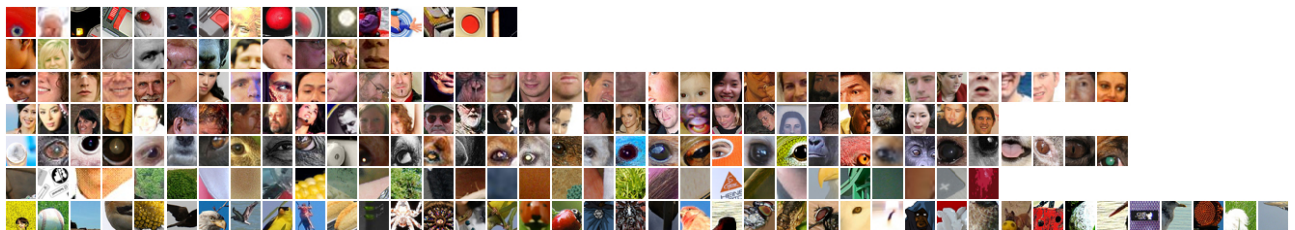


Fig. 18: Agrupació amb K-Means dels 200 *patches* més atesos al H10L6 utilitzant el criteri *Softmax*. Cada fila correspon a un clúster.