

Rețele neuronale și deep learning

Cursul 14: Modele multimodale și vizual-lingvistice; sinteza cursului



Prof. dr. Daniel Traian PELE

Academia de Studii Economice din București (ASE), Departamentul de Statistică și Econometrie
Academia Română, Institutul de Prognoză Economică (IPE)
IDA Institute Digital Assets ▪ AI4EFin ▪ MSCA Digital Finance

Obiectivele cursului

- Să explicați cum transformă un transformer pentru imagini o imagine în tokeni și unde se îndreaptă atenția sa, pe imagini satelitare reale.
- Să scrieți funcția de pierdere a preantrenării contrastive limbaj–imagine (CLIP), să o folosiți ca clasificator zero-shot și să o comparați cu o rețea neuronală convoluțională (CNN) antrenată pe imagini etichetate.
- Să descrieți cum leagă modelele vizual-lingvistice un codificator de imagini de un model de limbaj și să testați ce pot citi din grafice și tabele.
- Să sintetizați cursul: să alegeți modelul pentru o problemă de afaceri, de la regresia logistică la modelele fundație, cu evaluarea potrivită.

Abrevieri

ASE	Academia de Studii Economice din București
AUC	aria de sub curba caracteristică de operare (ROC)
BCE	Banca Centrală Europeană
CLIP	preantrenare contrastivă limbaj–imagine
CNN	rețea neuronală convoluțională
ESG	criterii de mediu, sociale și de guvernanță
GAN	rețea generativă adversarială
GNN	rețea neuronală pe grafuri
GPT	transformer generativ preantrenat
ICBE	Conferința internațională Business Excellence
InfoNCE	funcția de pierdere contrastivă InfoNCE (information noise-contrastive estimation)
IPE	Institutul de Prognoză Economică, Academia Română
LLaVA	asistentul lingvistic și vizual LLaVA (large language and vision assistant)
LLM	model lingvistic mare
LSTM	rețea cu memorie pe termen scurt și lung
MLP	perceptron multistrat
OCR	recunoașterea optică a caracterelor
RAG	generare augmentată prin regăsire
RGB	roșu, verde, albastru (canalele de culoare)
RL	reinforcement learning
ROC	caracteristica de operare a receptorului
SHAP	explicații aditive Shapley
VAE	autoencoder variațional
VaR	valoarea expusă la risc (Value-at-Risk)
ViT	transformer pentru imagini (vision transformer)
VLM	model vizual-lingvistic
XAI	inteligență artificială explicabilă

Cuprins

- Datele multimodale în afaceri
- Transformere pentru imagini
- CLIP: imagini și text în același spațiu
- Modele vizual-lingvistice: citirea graficelor și a tabelelor
- Sinteza cursului

Imagini, documente și grafice ca date economice

- **Imagini satelitare și aeriene**
  - Monitorizarea culturilor pentru creditul și asigurarea agricolă; utilizarea terenului pentru raportările de mediu, sociale și de guvernanță (ESG)
- **Documente**
  - Contracte scanate, facturi, rapoarte anuale: text, tabele și așezare în pagină, împreună
- **Grafice**
  - Analiștii citesc graficele de preț; le poate citi un model și are rost, cînd cifrele există?
- **Acest curs**
  - CLIP și un model vizual-lingvistic (VLM), evaluate pe imagini reale

(Radford et al., 2021; Bishop & Bishop, 2024)

Aplicația economică: utilizarea terenului pentru credit, asigurări și ESG

- **Creditul agricol**
  - O bancă verifică dacă parcelele finanțate sînt teren arabil și rămîn așa pe tot sezonul
- **Asigurări**
  - Daunele pentru inundații și secetă se compară cu imagini de dinainte și de după eveniment
- **Raportarea ESG**
  - Suprafața împădurită și zonele industriale din jurul activelor unui portofoliu
- **Problema datelor**
  - Imaginile sînt gratuite (Sentinel-2, la cîteva zile), etichetele sînt scumpe: de aici modelele zero-shot și cu puține exemple

(Helber et al., 2019)

Datele: imagini satelitare, grafice de piață, tabele ale curbei randamentelor

Date	Dimensiune	Sarcina
EuroSAT: imagini Sentinel-2 ale Europei, canalele roșu, verde, albastru (RGB)	27.000 de imagini. 64×64 pixeli. 10 clase	clasificarea utilizării terenului: zero-shot, cu puține exemple, antrenare completă
Prețurile zilnice S&P 500 (Yahoo Finance), desenate ca imagini	945 de ferestre de 120 de zile	regimul pieței din imagine față de cel din cifre
Curba randamentelor AAA din zona euro, Banca Centrală Europeană (BCE)	60 de tabele desenate ca imagini	poate un VLM să citească un număr dintr-un tabel?

- EuroSAT împărțit 80/20 stratificat (valoarea generatorului aleator: 20261108): 21 600 de imagini de antrenare și 5 400 de test.

(Helber et al., 2019)

Cuprins

- Datele multimodale în afaceri
- Transformere pentru imagini
- CLIP: imagini și text în același spațiu
- Modele vizual-lingvistice: citirea graficelor și a tabelelor
- Sinteza cursului

Un transformer pentru imagini citește imaginea ca o secvență de petice

- **Pe scurt:** imaginea se taie în petice de 16×16 pixeli, fiecare petic devine un vector, iar vectorii sînt tratați ca niște cuvinte.

$$\mathbf{z}_0 = \left[\mathbf{x}_{\text{cls}};\ \mathbf{E}\mathbf{p}_1;\ \ldots;\ \mathbf{E}\mathbf{p}_{196}\right] + \mathbf{E}_{\text{pos}}, \quad \mathbf{p}_j \in \mathbb{R}^{16\cdot 16\cdot 3}$$

$$\mathbf{z}_\ell = \mathbf{z}_{\ell-1} + \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{\ell-1})), \quad \mathbf{z}_\ell \leftarrow \mathbf{z}_\ell + \mathrm{MLP}(\mathrm{LN}(\mathbf{z}_\ell))$$

- **unde**
  - Imagine $224 \times 224 = 14 \times 14 = 196$ de petice + un token de clasă = 197 de tokeni
  - $\mathbf{E}$ = proiecția liniară a peticelor (lățime 768)
  - Atenție multi-cap cu 12 capete, 12 straturi
  - Normalizarea pe strat
  - Perceptron multistrat (MLP).

- **Transformerul pentru imagini (ViT)**: fără convoluție
  - Localitatea trebuie învățată din date.
- **De aici nevoia de preantrenare uriașă**: CLIP a văzut 400 de milioane de perechi imagine–text.

- **În practică:** cu puține imagini etichetate, un ViT preantrenat întrece cu mult un ViT antrenat de la zero.

(Dosovitskiy et al., 2021)

Exerciții practic: notebook §1

Petice și atenție pe imagini EuroSAT reale

- **Pe scurt:** sus: grila de 14×14 petice; jos: unde privește tokenul de clasă, mediat pe toate straturile.



Sursa: EuroSAT, imagini Sentinel-2 (Helber et al., 2019). Date: 4 imagini de test mărite la 224 x 224 de pixeli = 14 x 14 petice de 16 x 16; modelul CLIP ViT-B/16.

- Propagarea atenției: $\tilde{\mathbf{A}} = \prod_\ell (\frac{1}{2}\bar{\mathbf{A}}_\ell + \frac{1}{2}\mathbf{I})$, rîndul tokenului de clasă
  - $\bar{\mathbf{A}}_\ell$ = atenția stratului ℓ, mediată pe capete.

(Abnar & Zuidema, 2020)

NN_L14_vit_patches_attention

Cuprins

- Datele multimodale în afaceri
- Transformere pentru imagini
- CLIP: imagini și text în același spațiu
- Modele vizual-lingvistice: citirea graficelor și a tabelelor
- Sinteza cursului

CLIP învață potrivind imaginile cu descrierile lor

- **Pe scurt:** într-un lot de n perechi imagine–descriere, fiecare imagine își caută descrierea, și invers.

$$s_{ij} = \frac{\mathbf{u}_i^\top \mathbf{v}_j}{\tau}, \quad \mathbf{u}_i = \frac{f(I_i)}{\|f(I_i)\|}, \ \mathbf{v}_j = \frac{g(T_j)}{\|g(T_j)\|}$$

$$\mathcal{L} = -\frac{1}{2n} \sum_{i=1}^{n} \left[\log \frac{e^{s_{ii}}}{\sum_j e^{s_{ij}}} + \log \frac{e^{s_{ii}}}{\sum_j e^{s_{ji}}} \right]$$

- **unde**
  - f = codificatorul de imagini (ViT)
  - g = codificatorul de text
  - τ = temperatura învățată
  - Pierderea contrastivă (InfoNCE), în ambele direcții.

- **ViT-B/32**: 151,3 milioane de parametri
  - **ViT-B/16**: 149,6 milioane.
- **Nicio etichetă de clasă** nu a fost folosită la antrenare: doar descrieri de pe internet.

- **În practică:** un singur spațiu de embedding pentru imagini și text permite căutarea, etichetarea zero-shot și regăsirea pentru generarea augmentată prin regăsire (RAG).

(Radford et al., 2021; Oord et al., 2018)

Clasificarea zero-shot cu numele claselor ca prompturi textuale

- **Pe scurt:** se calculează embedding-ul unei propoziții pentru fiecare clasă și al imaginii, apoi se alege propoziția cea mai apropiată.

$$\mathbf{w}_c = \operatorname{norm}\left(\frac{1}{M}\sum_{m=1}^{M} g\left(\text{template}_m(c)\right)\right)$$

$$\hat{c}(I) = \arg\max_c \ \mathbf{u}(I)^\top \mathbf{w}_c$$

- **Cel mai bun zero-shot**: CLIP ViT-B/16, 56,4 % pe 5 400 de imagini de test, fără niciun exemplu etichetat.
- **Nivelul hazardului** cu 10 clase: 10 %.

- **unde**
  - Template_m = o propoziție ca „a satellite photo of {class}."
  - $M = 1$ sau 5 șabloane (ansamblu de prompturi)
  - Descrierile claselor ca la Radford et al. (2021).

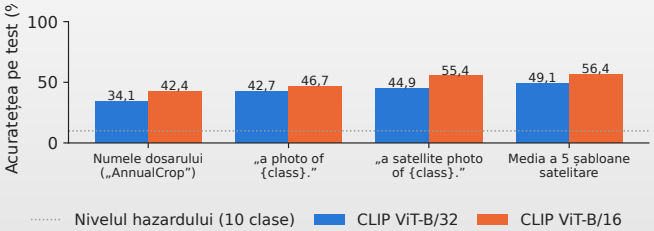
- **În practică:** vectorii de ponderi sînt scriși în limba engleză, deci un expert al domeniului poate modifica clasificatorul fără reantrenare.

(Radford et al., 2021)

Exercițiu practic: notebook §2

Cuvintele din prompt schimbă acuratețea zero-shot

- **Pe scurt:** a-i spune lui CLIP că imaginea este o fotografie satelitară valorează mai mult decît un model mai mare.



Sursa: EuroSAT, imagini Sentinel-2 (Helber et al., 2019). Date: 5400 de imagini de test, fără antrenare.

- ViT-B/16: numele dosarului 42,4 %, „a satellite photo of” 55,4 %, ansamblul de 5 șabloane 56,4 %
  - ViT-B/32 ansamblu 49,1 %.

- **În practică:** formularea promptului este un hiperparametru: alegeți-o pe imagini de validare, apoi fixați-o.

NN_L14_zeroshot_prompts

Predicțiile zero-shot pe opt imagini de test alese aleator

- **Pe scurt:** verde: corect; roșu: greșit; procentul este probabilitatea softmax a clasei alese.



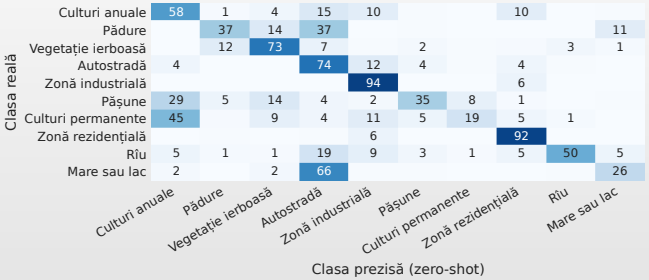
Sursa: EuroSAT, imagini Sentinel-2 (Helber et al., 2019). Date: 8 imagini de test alese aleator (valoare fixă); verde = corect, roșu = greșit.

- 5 din 8 corecte în acest eșantion.

NN_L14_zeroshot_examples

Erorile zero-shot la utilizări ale terenului asemănătoare vizual

- **Pe scurt:** rîndurile: clasa reală; coloanele: clasa prezisă; fiecare rînd însumează 100 %.



Sursa: EuroSAT, imagini Sentinel-2 (Helber et al., 2019). Date: 5400 de imagini de test; procente pe rînd.

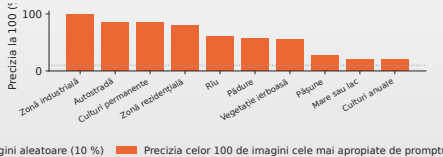
- Cea mai bună clasă: zonă industrială (93,6 %)
  - Cea mai slabă: culturi permanente (18,6 %)
  - Cea mai mare confuzie: 399 din 600 imagini „mare sau lac” clasificate drept „autostradă”.

NN_L14_zeroshot_confusion

Căutarea imaginilor după text cu aceleași embedding-uri

- **Pe scurt:** ordonăm toate imaginile de test după similaritatea cu o propoziție; o bancă poate găsi terenuri inundate sau clădiri noi fără etichete.

$$\mathrm{P@}k(c) = \frac{1}{k} \sum_{j \in \mathrm{top}_k(\mathbf{w}_c)} \mathbb{1}\{y_j = c\}, \quad \mathrm{top}_k(\mathbf{w}_c) = \text{cele } k \text{ imagini cu cel mai mare } \mathbf{u}(I_j)^\top \mathbf{w}_c$$



Sursa: EuroSAT, imagini Sentinel-2 (Helber et al., 2019). Date: 5400 de imagini de test; modelul CLIP ViT-B/16, 5 șabloane.

- Precizia medie a primelor 100: 58,9 %
  - Cea mai bună: zonă industrială (99,0 %), cea mai slabă: culturi anuale (20,0 %).

- **În practică:** căutați întîi, etichetați rezultatele, apoi antrenați un clasificator: cea mai ieftină cale spre un set de imagini etichetate.

NN_L14_text_search

Un clasificator liniar pe trăsăturile CLIP: embedding plus regresie logistică

- **Pe scurt:** CLIP rămîne fix și se antrenează doar o regresie logistică pe embedding-urile imaginilor.

$$\hat{p}(c \mid I) = \operatorname{softmax}\left(\mathbf{W}\mathbf{u}(I) + \mathbf{b}\right)_c$$

$$\min_{\mathbf{W},\mathbf{b}} \; -\sum_i \log \hat{p}(c_i \mid I_i) + \frac{1}{2C}\|\mathbf{W}\|_F^2$$

- **Toate imaginile de antrenare**: 95,5 % (ViT-B/32: 94,8 %).
- **Depășește zero-shot** deja cu o singură imagine etichetată pe clasă: 59,0 %.

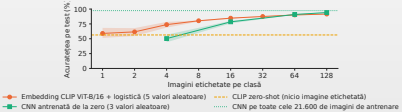
- **unde**
  - $\mathbf{u}(I)$ = embedding-ul normalizat ViT-B/16 (512 de dimensiuni)
  - $C = 10$, fixat a priori
  - k imagini etichetate pe clasă, 5 rulări.

- **În practică:** un astfel de clasificator se antrenează în cîteva secunde pe laptop: primul lucru de încercat cînd aveți cîteva sute de etichete.

Exercițiu practic: notebook §3

Trăsăturile preantrenate și o CNN antrenată de la zero, după numărul de etichete

- **Pe scurt:** o imagine etichetată pe clasă depășește deja zero-shot; de la circa 64 pe clasă, o rețea mică antrenată de la zero o ajunge din urmă.



Sursa: EuroSAT, imagini Sentinel-2 (Helber et al., 2019). Date: 5400 de imagini de test; zona colorată = intervalul pe valorile generatorului aleator.

- 16 imagini pe clasă: clasificatorul pe embedding 84,7 %, CNN 78,5 %
  - 128 pe clasă: 91,6 % față de 94,1 %
  - Toate cele 21 600 de imagini: CNN 97,4 %, clasificatorul pe embedding 95,5 %.

- **În practică:** cu multe etichete, o CNN mică recuperează
  - Cu puține etichete, porniți de la un model fundație.

NN_L14_fewshot_eurosat

Dimensiunea și viteza celor trei modele de imagini

Model	Parametri	ms pe imagine	Acuratețea pe test (%)
CNN de la zero (toate etichetele)	289,2 mii	–	97,4
CLIP ViT-B/32 + clasificator	151,3 mil.	4,23	94,8
CLIP ViT-B/16 + clasificator	149,6 mil.	20,49	95,5
Qwen2-VL-2B (întrebări despre grafice)	2,21 mld.	262	–

- Timp măsurat pe un laptop (procesor grafic Apple silicon), pentru calculul embedding-ului unei imagini în loturi; timpul VLM include întrebarea.
- Un petic mai mare (32 de pixeli în loc de 16) înseamnă de 4 ori mai puțini tokeni: mai rapid, puțin mai puțin precis.

(calcul propriu)

Cuprins

- Datele multimodale în afaceri
- Transformere pentru imagini
- CLIP: imagini și text în același spațiu
- Modele vizual-lingvistice: citirea graficelor și a tabelelor
- Sinteza cursului

Un VLM: un codificator de imagini conectat la un model de limbaj

- **Pe scurt:** peticele imaginii sînt proiectate în spațiul embedding-urilor de cuvinte al unui model lingvistic mare (LLM), care le citește ca tokeni suplimentari.

$$\mathbf{H}_v = \mathbf{W}_p\, f_{\text{ViT}}(I)$$
$$p(y_t \mid y_{<t}, I, q) = \text{LLM}_\theta\big([\mathbf{H}_v;\ \mathbf{e}(q);\ \mathbf{e}(y_{<t})]\big)$$

- **unde**
  - f_{ViT} = codificatorul vizual
  - $\mathbf{W}_p$ = proiectorul (un MLP)
  - q = întrebarea
  - Antrenat întîi pe perechi imagine–descriere, apoi pe instrucțiuni vizuale.

- **Asistentul lingvistic și vizual (LLaVA)** a introdus această rețetă
  - Qwen2-VL adaugă imagini de orice rezoluție.
- **Modelul nostru**: Qwen2-VL-2B-Instruct, 2,21 miliarde de parametri, rulat local.

- **În practică:** răspunsul este text generat: restrîngeți-l (da/nu, un număr) și evaluați-l ca pe un clasificator.

(Liu et al., 2023; Wang et al., 2024)

Regimurile pieței ca problemă de recunoaștere vizuală

- **Pe scurt:** desenăm ultimele 120 de zile ale S&P 500 ca imagine și întrebăm dacă piața este în criză, cum ar face un analist tehnic.

$$S_t = \mathbb{1}\left\{\sigma_t^{(20)} > Q_{0.9}\left(\sigma_{t-750:t-1}^{(20)}\right)\right\}$$
$$\sigma_t^{(20)} = \sqrt{252}\,\mathrm{sd}(r_{t-19:t})$$

- **unde**
  - S_t = eticheta de criză la sfîrșitul ferestrei: volatilitatea pe 20 de zile peste percentila 90 glisantă din ultimii trei ani (doar date din trecut)
  - Ferestre la fiecare 5 zile
  - Antrenare pînă în 2018, test pe 394 de ferestre din 2019.

- **Patru metode pe aceleași ferestre**
  - Cifre + regresie logistică (randamente, volatilitate, scădere de la maxim)
  - Embedding-ul CLIP al imaginii + regresie logistică
  - CLIP zero-shot: „calm, în creștere constantă" față de „volatil, prăbușire bruscă"
  - Qwen2-VL zero-shot: „este piața agitată?" da sau nu

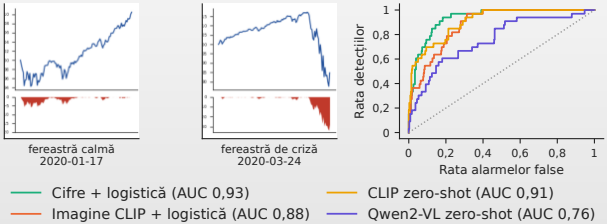
- **În practică:** o etichetă bazată pe cuantile glisante evită etichetarea manuală și folosirea viitorului.

(Jheng, Pele, Tak & Găman, 2026)

Exercițiu practic: notebook §4

Detectarea crizelor de piață din grafice și din cifre

- **Pe scurt:** stînga: o fereastră calmă și una de criză, așa cum le văd modelele; dreapta: curbele caracteristicii de operare a receptorului (ROC) pe anii de test.



Sursa: Yahoo Finance, cotații zilnice. Date: S&P 500, ferestre de 120 de zile de tranzacționare la fiecare 5 zile; test 7 ianuarie 2019 - 30 octombrie 2026.

- Aria de sub curba ROC (AUC): cifre 0,933, clasificator pe CLIP 0,880, CLIP zero-shot 0,907, Qwen2-VL zero-shot 0,758
  - Ponderea crizelor în test 8,4 %.

NN_L14_chart_regime

Detectarea crizelor: capacitatea de separare și precizia pe anii de test

Metoda	AUC	Prec. medie	Etichete
Cifre + logistică	0,933	0,578	da
Imagine CLIP + logistică	0,880	0,449	da
CLIP zero-shot	0,907	0,605	nu
Qwen2-VL zero-shot	0,758	0,302	nu

- **Cel mai bun**: cifrele.
- **Precizia medie** a unui scor aleator = ponderea crizelor, 8,4 %.
- **Lecția**: cînd cifrele există, imaginea adaugă un pas cu pierdere de informație
  - Imaginile își au locul cînd cifrele lipsesc (scanări, capturi de ecran, imagini satelitare).

(calcul propriu, după Jheng, Pele, Tak & Găman, 2026)

Exercițiu practic: notebook §4

Citirea tendințelor și a celulelor de tabel cu un VLM de 2 miliarde de parametri

- **Întrebarea despre tendință**, pe 394 de grafice de test: „prețul a încheiat mai sus sau mai jos?"
  - Acuratețe 57,1 % la pragul de 0,5, sub răspunsul constant cu clasa majoritară (74,9 %)
  - Totuși AUC 0,974: ordonarea este corectă, pragul este deplasat — recalibrați pe grafice de validare
- **Întrebarea despre tabel**, pe 60 de tabele ale curbei randamentelor BCE: „care este randamentul la 10 ani?"
  - Răspunsuri exacte 25,0 %; eroarea absolută medie 0,192 puncte procentuale; numere extrase 100,0 %
- **De ce contează**
  - Citirea unei celule înseamnă recunoaștere optică a caracterelor (OCR) plus găsirea coloanei: erorile sînt tăcute și plauzibile
  - Pentru documente cu structură cunoscută, extrageți cifrele cu un program de citire și folosiți VLM doar unde structura variază

(calcul propriu; Wang et al., 2024)

Exercițiu practic: notebook §5

Rolul unui VLM în fluxul documentelor unei bănci

- **1. Preluarea**
  - Pagină scanată sau fișier electronic; șabloanele cunoscute merg la un program de citire bazat pe reguli
- **2. Citirea**
  - OCR și așezarea în pagină pentru text și tabele; un VLM pentru pagini libere, grafice și ștampile
- **3. Verificarea**
  - Totalurile trebuie să se adune, datele să fie valide, valorile în intervale plauzibile; o verificare eșuată merge la un om
- **4. Măsurarea**
  - Rata de potrivire exactă pe fiecare cîmp, pe un eșantion etichetat, ca la tabelele de randamente (aici 25,0 %)

Capacitățile și limitele modelelor multimodale

- **Pot**
  - Eticheta imagini fără date de antrenare (56,4 % pe EuroSAT)
  - Oferi trăsături care au nevoie de puține etichete (84,7 % cu 16 pe clasă)
  - Căuta imagini după text și text după imagini
  - Răspunde la întrebări simple despre grafice și documente
- **Nu pot (în mod fiabil)**
  - Înlocui cifrele din spatele unui grafic
  - Citi corect fiecare celulă a unui tabel (aici, 25,0 % exact)
  - Explica fidel de ce au răspuns astfel
  - Separa fără exemple clasele care arată la fel
- Măsuri de control: un set de test etichetat, răspunsuri restrînse, verificarea umană a cazurilor cu încredere scăzută, jurnalizarea fiecărui prompt și a fiecărei imagini.

Cuprins

- Datele multimodale în afaceri
- Transformere pentru imagini
- CLIP: imagini și text în același spațiu
- Modele vizual-lingvistice: citirea graficelor și a tabelelor
- Sinteza cursului

Metodele cursului după tipul de date

Date	Modele	Cursuri și aplicații
Tabele	regresie logistică, MLP, regularizare	1–3: scoring de credit, fraudă
Imagini	CNN, transferul învățării, ViT, CLIP	4, 14: utilizarea terenului, documente
Secvențe	rețele cu memorie pe termen scurt și lung (LSTM), rețele cu expansiune în baze, modele fundație	5, 13: prețuri, consum, valoarea expusă la risc (VaR)
Text	embedding, transformere, LLM, generare augmentată prin regăsire (RAG)	6, 7, 12: sentiment, bănci centrale
Generare	autoencoder variațional (VAE), rețea generativă adversarială (GAN), difuzie	8: date sintetice, anomalii
Grafuri	rețea neuronală pe grafuri (GNN)	9: rețele de tranzacții
Decizii	reinforcement learning (RL)	10: tranzacționare, acoperirea riscului, baterii
Încredere	inteligență artificială explicabilă (XAI): explicații aditive Shapley (SHAP), gradienți	11: auditul modelelor de credit

Alegerea modelului pentru o problemă de afaceri

- **1. Porniți de la decizie și de la costul erorilor**
  - O fraudă ratată și o alarmă falsă nu costă la fel; alegeți metrica înaintea modelului
- **2. Estimați întîi reperul**
  - Regresie logistică, prognoza naivă, o autoregresie Lasso, un model de volatilitate, frecvențe de cuvinte cu un clasificator liniar: cîteva minute, ușor de interpretat
- **3. Numărați etichetele**
  - Puține etichete: model fundație zero-shot sau cu un clasificator pe embedding; multe etichete: o rețea antrenată
- **4. Potriviți arhitectura cu datele**
  - Tabele: MLP sau boosting; imagini: CNN sau ViT; secvențe: rețele recurente sau transformer; grafuri: GNN
- **5. Demonstrați**
  - Testare o singură dată, mai multe rulări, un test de semnificație, calibrare, o explicație pentru autoritatea de supraveghere

Cinci reguli de evaluare aplicate pe parcursul cursului

- **Fără scurgeri de informație**: scalările, prompturile și pragurile se aleg pe datele de antrenare sau de validare
  - Seriile de timp se împart cronologic.
- **Un reper lîngă fiecare rețea**: aici, o CNN antrenată de la zero lîngă CLIP, cifrele lîngă imaginile graficelor.
- **Rulări repetate și intervale**: rezultatele cu puține exemple de mai sus sînt medii pe 5 rulări, cu amplitudinea lor.
- **Metrici pentru evenimente rare**: precizia medie și AUC pentru crize, nu acuratețea.
- **Date reale, rezultate reproductibile**: fiecare cifră provine dintr-un program și dintr-un set de date public, cu un Quantlet.

Modelele fundație pentru imagini, text și serii de timp

- **Tiparul**
  - Preantrenare o singură dată pe volume uriașe de date neetichetate; utilizare zero-shot, cu cîteva exemple sau cu un mic clasificator antrenat
- **Text** (cursul 12)
  - Prompturi și embedding LLM pentru sentiment, RAG pe discursurile BCE, LLM-VaR
- **Serii de timp** (cursul 13)
  - Modele de tip Chronos prognozează prețuri, consum și VaR fără antrenare
- **Imagini** (acest curs)
  - CLIP zero-shot 56,4 %; clasificator pe toate etichetele 95,5 %
- **Avertismentul comun**
  - Datele de preantrenare pot conține setul dumneavoastră de test; evaluați pe date pe care modelul nu le-a putut vedea

Zece formule ale cursului (1/2): învățarea

Idee	Formula	Cursul
Neuronul logistic	$\hat{p} = \sigma(\mathbf{w}^\top \mathbf{x} + b)$	1
Pierderea entropie încrucișată	$\mathcal{L} = -\sum_i [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)]$	1–2
Propagarea înapoi	$\boldsymbol{\delta}^{(\ell)} = (\mathbf{W}^{(\ell+1)\top} \boldsymbol{\delta}^{(\ell+1)}) \odot f'(\mathbf{z}^{(\ell)})$	2
Pasul de gradient (Adam fără momente)	$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}$	2
Penalizarea ponderilor	$\mathcal{L} + \lambda \lVert \boldsymbol{\theta} \rVert_2^2$	3

⊡ Toate simbolurile, ca în cursurile citate; σ = funcția logistică, $\odot$ = produsul element cu element.

Zece formule ale cursului (2/2): arhitecturi

Idee	Formula	Cursul
Convoluția	$(\mathbf{K} * \mathbf{X})_{ij} = \sum_{u,v} K_{uv} X_{i+u,j+v}$	4
Starea recurentă	$\mathbf{h}_t = \tanh(\mathbf{W}\mathbf{h}_{t-1} + \mathbf{U}\mathbf{x}_t + \mathbf{b})$	5
Atenția	$\mathrm{softmax}(\mathbf{Q}\mathbf{K}^\top / \sqrt{d_k})\mathbf{V}$	7, 12–14
Convoluția pe graf	$\mathbf{H}' = \sigma(\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}\mathbf{H}\mathbf{W})$	9
Pierderea contrastivă imagine–text	$-\log \frac{e^{s_{ii}}}{\sum_j e^{s_{ij}}}$	14

⊡ $\mathbf{K}$ = nucleul; $\tilde{\mathbf{A}}$ = matricea de adiacență cu bucle, $\tilde{\mathbf{D}}$ matricea gradelor; s_{ij} = similaritatea scalată imagine–text.

Utilizarea responsabilă a modelelor din curs

- **Reglementare**
  - Regulamentul UE privind inteligența artificială (2024/1689) include scoringul de credit printre utilizările cu risc ridicat: documentație, supraveghere umană, guvernanța datelor
- **Documentare**
  - O fișă a modelului: date, perioadă, metrică, reper, rulări, cazuri cunoscute de eșec (aici: mare sau lac citit drept autostradă)
- **Monitorizare**
  - Retestare pe date noi: regimuri de preț, noi sezoane satelitare, teme noi în discursuri
- **Confidențialitate**
  - Modelele deschise rulate local păstrează documentele și imaginile clienților în instituție

Cerințele unui proiect bun pentru acest curs

- **Întrebare și date**
  - O decizie de afaceri, un set de date public, perioada și licența lui
- **Întîi reperul**
  - Un model liniar sau naiv, raportat lîngă fiecare rețea
- **Verificarea scurgerilor**
  - Împărțire în ordine cronologică, scalări estimate pe antrenare, testul folosit o singură dată
- **Incertitudine**
  - Mai multe rulări, intervale bootstrap sau un test Diebold–Mariano
- **Explicație**
  - SHAP sau gradienți pentru decizia care contează
- **Reproductibil**
  - Un singur program, de la datele brute la fiecare cifră și grafic

Probleme de cercetare deschise pentru lucrările de disertație

- **Calibrarea în prezența derivei**: benzi conformale care rămîn valide cînd pieţele îşi schimbă regimul.
- **Evaluarea fără scurgeri de informație** a modelelor fundație antrenate pe internet.
- **Explicații stabile**: rezistă valorile SHAP ale unui model de credit la o reantrenare?
- **Modele mici pentru instituții mici**: cînd ajunge un model de 2–3 miliarde de parametri pe un laptop?
- **Semnale de risc multimodale**: imagini satelitare, text și prețuri într-un singur model de avertizare timpurie.
- **Date românești**: modele de limbaj și prognoze pentru textul, pieţele și energia din România.

Test 1: clasificatorul zero-shot

- O bancă vrea să marcheze terenurile inundate în imagini satelitare, dar nu are nicio imagine etichetată
  - Ce poate face cu CLIP?

Răspuns

- Scrie prompturi pentru clase („a satellite photo of a flooded field", „... of a dry field"), le calculează embedding-urile și clasifică după similaritatea cosinus
  - Apoi etichetează cîteva sute de imagini pentru validare și pentru un clasificator pe embedding.

Test 2: imagini sau cifre

- De ce a detectat modelul logistic pe cifre crizele de piață cel puțin la fel de bine ca modelele care citesc imaginea graficului?

Răspuns

- Imaginea este desenată din aceleași cifre, cu pierdere de rezoluție
  - Cifrele conțin direct volatilitatea.
  - Imaginile ajută cînd nu există cifre.

Test 3: etichetele

- Cu 16 imagini etichetate pe clasă, ce alegeți: o CNN de la zero sau un clasificator pe embedding-urile CLIP?

Răspuns

- Clasificatorul pe embedding: 84,7 %, față de 78,5 % pentru CNN, pe datele noastre.

Recapitulare

- Un ViT citește imaginile ca tokeni-petic
  - CLIP le aliniază cu textul printr-o pierdere contrastivă.
- CLIP zero-shot etichetează imaginile satelitare cu 56,4 %, fără etichete
  - Un clasificator pe embedding ajunge la 95,5 %.
- Modelele VLM răspund la întrebări despre grafice și tabele, dar trebuie evaluate ca orice clasificator.
- Cursul într-un rînd: întîi reperul, date reale, fără scurgeri de informație, mai multe rulări, un test, o explicație.

Cercetările grupului nostru

- **Detectarea anomaliilor cu modele vizual-lingvistice** (premiul pentru cea mai bună lucrare, Conferința internațională Business Excellence (ICBE) 2026): indicatorii de piață sînt proiectați pe imagini cu mai multe panouri, iar un VLM (GPT-4.1, transformer generativ preantrenat) clasifică regimul zero-shot, cu etichete bazate pe cuantile glisante, și își explică alegerea
  - Validare pe date noi pentru S&P 500 (Jheng, Pele, Tak & Găman, 2026, PICBE).
- **Detectarea explicabilă a fraudei**: modele tabulare și rețele pe grafuri, combinate cu SHAP și ablația grafului (Diaconu & Pele, 2026, PICBE) — studiul de caz 1.
- **LLM-VaR și LLM-ES**: modelele de limbaj ca modele de risc zero-shot (Pele et al., 2026, Expert Systems with Applications) — cursurile 12 și 13.
- **Portofolii care țin cont de ESG**: selecția activelor după mai multe criterii, cu scoruri ESG (Găman, Tak, Jheng & Pele, 2026, PICBE).
- Rezultatele sînt preluate din rezumatele lucrărilor; cifrele de pe celelalte slide-uri provin din datele noastre.

(Academia de Studii Economice din București, ASE; Academia Română, Institutul de Prognoză Economică, IPE)

Lecturi și resurse suplimentare

- **OpenCLIP**, modele și evaluări: github.com/mlfoundations/open_clip.
- **Qwen2-VL**: github.com/QwenLM/Qwen2-VL.
- **Bishop & Bishop (2024)**, capitolele 12 și 20: bishopbook.com.
- **Quantinar**: Digital Transformation on Finance: quantinar.com.
- **EuroSAT**, date: github.com/phelber/EuroSAT.
- **Acest curs**: notebook, seminar și Quantlets — github.com/danpele/NEURAL_BIZ.

Bibliografie (1/2)

- Abnar, S. & Zuidema, W. (2020). Quantifying attention flow in transformers. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4190–4197. doi:10.18653/v1/2020.acl-main.385
- Bishop, C. M. & Bishop, H. (2024). *Deep Learning: Foundations and Concepts*. Springer. doi:10.1007/978-3-031-45468-4
- Diaconu, D. & Pele, D. T. (2026). Explainable AI in financial fraud detection: Evidence from the IEEE-CIS fraud dataset. *Proceedings of the International Conference on Business Excellence*, 20(1), 118–132. doi:10.2478/picbe-2026-0015
- Dosovitskiy, A. et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*. arXiv:2010.11929
- Găman, Ș., Tak, R., Jheng, S.-L. & Pele, D. T. (2026). ESG-aware asset selection and portfolio optimization: A comparative study of multi-criteria decision-making methods. *Proceedings of the International Conference on Business Excellence*, 20(1), 133–147. doi:10.2478/picbe-2026-0016
- Helber, P., Bischke, B., Dengel, A. & Borth, D. (2019). EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7), 2217–2226. doi:10.1109/JSTARS.2019.2918242
- Jheng, S.-L., Pele, D. T., Tak, R. & Găman, Ș. (2026). A multimodal vision-language framework for financial anomaly detection. *Proceedings of the International Conference on Business Excellence*, 20(1), 191–201. doi:10.2478/picbe-2026-0020

Bibliografie (2/2)

- Liu, H., Li, C., Wu, Q. & Lee, Y. J. (2023). Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36. arXiv:2304.08485
- Oord, A. van den, Li, Y. & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv:1807.03748
- Pele, D. T., Bolovăneanu, V., Lin, M.-B., Ren, R., Ginavar, A. T., Spilak, B., Andrei, A.-V., Toma, F.-M., Lessmann, S. & Härdle, W. K. (2026). In the beginning was the Word: LLM-VaR and LLM-ES. *Expert Systems with Applications*, 295, 128676. doi:10.1016/j.eswa.2025.128676
- Radford, A. et al. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763. arXiv:2103.00020
- Wang, P. et al. (2024). Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. arXiv:2409.12191

Anexa A: pierderea contrastivă ca problemă de clasificare

- Rîndul i al matricei $\mathbf{S} = (s_{ij})$ este un clasificator softmax peste n descrieri, cu clasa corectă i: pierderea este entropia încrucișată cu etichetele $0, \ldots, n-1$.
- Un lot mai mare înseamnă mai multe exemple negative și o sarcină mai grea
  - CLIP a folosit $n = 32.768$.
- Clasificarea zero-shot este același calcul, cu cele n descrieri înlocuite de cele C prompturi ale claselor.

(Radford et al., 2021; Oord et al., 2018)

Anexa B: detalii de implementare

- CLIP: openai/clip-vit-base-patch32 și -patch16 (Hugging Face transformers), imagini redimensionate la 224 de pixeli
  - Regresia logistică din scikit-learn.
- CNN: trei blocuri convoluționale (32, 64, 128 de canale), AdamW cu program one-cycle, oglindiri și rotații ca augmentare
  - 289,2 mii de parametri.
- VLM: Qwen/Qwen2-VL-2B-Instruct, float16
  - Răspunsurile da/nu evaluate prin probabilitățile primului token
  - Numerele prin decodare lacomă.
- Valoarea generatorului aleator: 20261108
  - Analiza completă a durat 0,6 minute la ultima rulare (embedding-urile și răspunsurile stocate).