



Recognizing Front Pages of Historical Newspapers: From Deep Learning to Automating Digitization Workflow

Tan Lu, KBR Data Science Lab & VUB WIDS

KBR  Where time
is treasured

 **VUB**
VRIJE
UNIVERSITEIT
BRUSSEL

Background: The BelgicaPress Project



- The **BelgicaPress** project is an (ongoing) large-scale digitization project managed by the DIGIT and ICT departments of KBR.
- Currently grants access to **138** Belgian newspapers known as "major press", mostly daily newspapers (**1814 - 1987**). With **full-text search** enabled for **4,138,188** pages.
- The priority period extends from the Belgian Revolution and the creation of independent Belgium (1830) to 31.12.1950. Including most of the newspapers published under German censorship during the two world wars.
- Unlimited access for scientific research or teaching illustration purposes.

Background: The BelgicaPress Project



- Organize the raw digitization outputs for proper online publication: recognizing **front pages**.
- From manual checking to AI-assisted automation
 - OCR -> the presence of a newspaper title + layout analysis
 - OCR outputs are potentially noisy
 - variance in layout style complicates the recognition of title region using heuristics

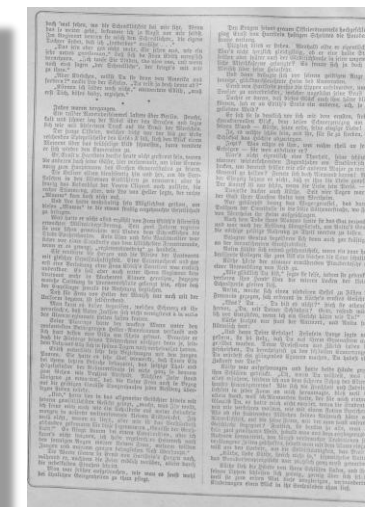
Challenge for OCR based approach: layout complexities

Den Vaderlander
The Patriot (Dutch)
April 1939



Echo De La Bourse (L')
The Stock Exchange Echo
(French)
February 1942

Avenir Du Luxembourg (L')
Future of Luxembourg
(French)
December 1902



Die Fliegende Taube
(Beilage)
The Flying Dove
(German),
January 1900

Background: The BelgicaPress Project



- Organize the raw digitization outputs for proper online publication: recognizing **front pages**.
- From manual checking to AI-assisted automation
 - OCR -> the presence of a newspaper title + layout analysis
 - OCR outputs are potentially noisy
 - variance in layout style complicates the recognition of title region using heuristics
 - ✓ Deep learning -> end-to-end framework using neural networks

Challenge for AI based approach: visual similarities

Cross-similarities between front and non-front pages of different newspaper titles: examples with Avenir Du Luxembourg (Future of Luxembourg)



1902.12.11 – page 4



1936.02.02 – page 4



1936.02.02 – page 5



1936.02.02 – page 7



1902.12.11 – page 1



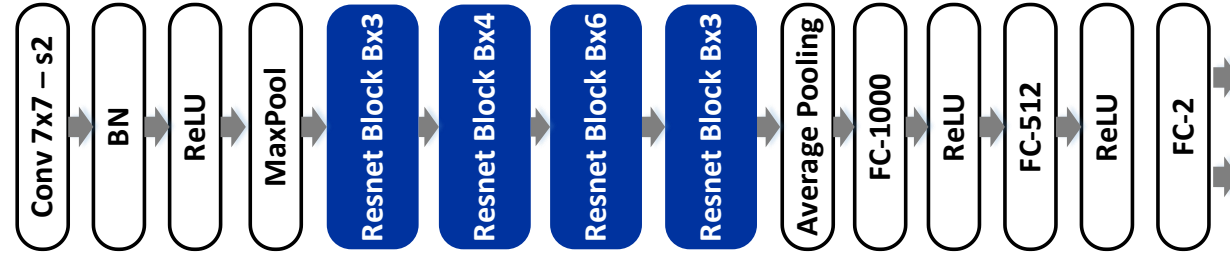
1936.02.02 – page 1



Front pages from other newspapers

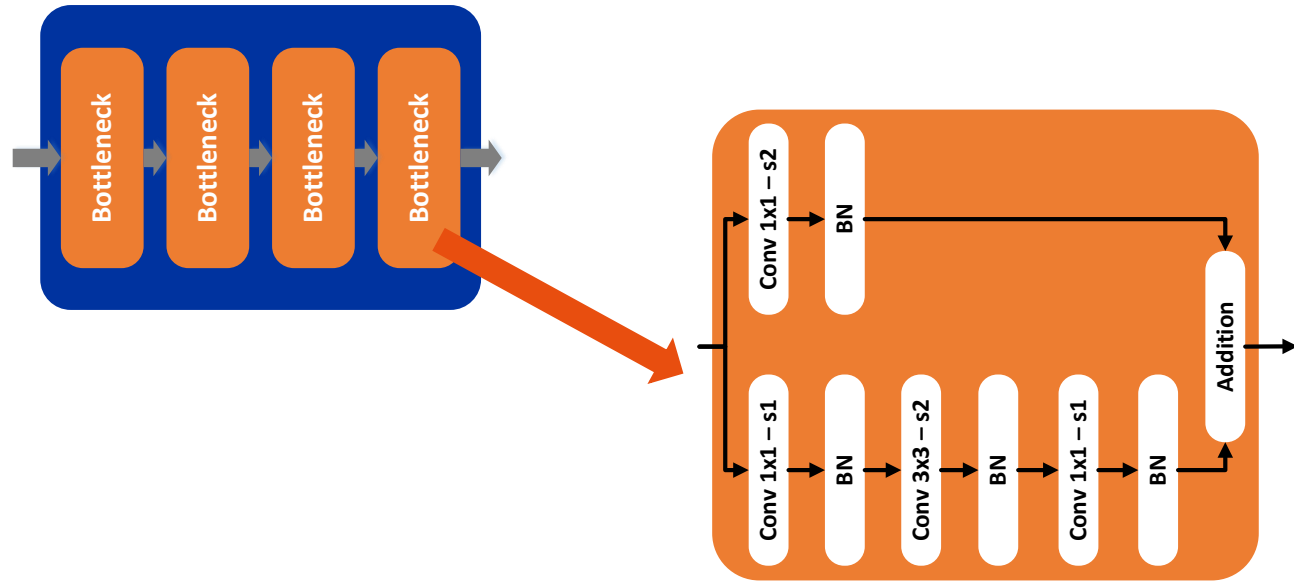


AI Model Backbone: Resnet



Front Page

Not Front Page



Preliminary Investigation

Title	Configuration					Performance				MISTAKES
	Train/Test	Year	Month	Pages		Precision	Recall	bAccuracy	F1	
				Total	Front					
Vooruit (15981364)	Training	1923	1, 2, 5, 6, 7, 8, 9, 10, 11, 12	1656	258	N.A.				1923-04-16-Page1 1923-04-16-Page3
	Testing	1923	3,4	406	53	0.98	0.98	0.99	0.98	
		1924	8	160	27	1	1	1	1	
		1923	1 ~ 12	2062	311	0.997	0.997	0.998	0.997	
Le XX Siecle (16343920)	Training	1923	1, 3, 4, 5, 6, 7, 9, 10, 11, 12	1820	295	N.A.				
	Testing	1923	2,8	344	58	1	1	1	1	
		1924	3	196	31	1	1	1	1	
		1923	1 ~ 12	2164	353	1	1	1	1	
Journal de Bruxelles (16411586)	Training	1923	1, 2, 3, 4, 5, 7, 8, 9, 10, 11	1180	295	N.A.				
	Testing	1923	6,12	236	59	1	1	1	1	
		1924	2	112	28	1	1	1	1	
		1923	1 ~ 12	1416	354	1	1	1	1	
Het Nieuws van den Dag (17159617)	Training	1923	1, 2, 3, 4, 5, 6, 8, 10, 11, 12	1572	255	N.A.				
	Testing	1923	7,9	332	52	1	1	1	1	
		1924	4	138	26	1	1	1	1	
		1923	1 ~ 12	1904	367	1	1	1	1	
Vlaanderen (17164874)	Training	1923	1, 2, 4, 5, 6, 7, 8, 9, 10, 12	344	43	N.A.				
	Testing	1923	3,11	64	8	1	1	1	1	
		1924	7	32	4	1	1	1	1	
		1923	1 ~ 12	408	51	1	1	1	1	
De Volksgazet (17165064)	Training	1923	1, 2, 3, 4, 5, 6, 7, 9, 11, 12	1246	260	N.A.				1923-08-06-Pages1
	Testing	1923	8,10	258	54	0.98	0.98	0.99	0.98	
		1924	6	134	25	1	1	1	1	
		1923	1 ~ 12	1504	314	1	0.997	0.998	0.998	
De Standaard (17172097)	Training	1923	1, 2, 3, 4, 7, 8, 9, 10, 11, 12	1552	298	N.A.				1923-05-03-Pages3 1923-07-29-Pages1 1923-08-11-Pages1
	Testing	1923	5,6	318	59	0.98	1	1	0.99	
		1924	5	168	30	1	1	1	1	
		1923	1 ~ 12	1870	357	0.997	0.994	0.996	0.996	
Independence Belge (17229173)	Training	1923	1, 2, 3, 4, 5, 6, 7, 8, 10, 12	1564	296	N.A.				1923-05-30-Pages5
	Testing	1923	9,11	314	57	1	1	1	1	
		1924	1	162	30	1	1	1	1	
		1923	1 ~ 12	1878	353	0.997	1	1	0.999	

Configuration:

- ResNeSt (resnest50_fast_1s4x24d), parameter size ~28M, basic CE loss, without hyperparameters optimization.
- Prioritize on classification performance, input images are scaled to 1024x1024.
- 8 different newspaper titles over the period of one year (80% training and 20% evaluation).
- ~10K pages, ~2K front pages.

Preliminary Investigation

Mistakes by AI, or, by human?

Title	Configuration					Performance				MISTAKES
	Train/Test	Year	Month	Pages		Precision	Recall	bAccuracy	F1	
				Total	Front					
Vooruit (15981364)	Training	1923	1, 2, 5, 6, 7, 8, 9, 10, 11, 12	1656	258	N.A.				1923-04-16-Page1 1923-04-16-Page3
	Testing	1923	3,4	406	53	0.98	0.98	0.99	0.98	
		1924	8	160	27	1	1	1	1	
		1923	1 ~ 12	2062	311	0.997	0.997	0.998	0.997	
Le XX Siecle (16343920)	Training	1923	1, 3, 4, 5, 6, 7, 9, 10, 11, 12	1820	295	N.A.				
	Testing	1923	2,8	344	58	1	1	1	1	
		1924	3	196	31	1	1	1	1	
		1923	1 ~ 12	2164	353	1	1	1	1	
Journal de Bruxelles (16411586)	Training	1923	1, 2, 3, 4, 5, 7, 8, 9, 10, 11	1180	295	N.A.				
	Testing	1923	6,12	236	59	1	1	1	1	
		1924	2	112	28	1	1	1	1	
		1923	1 ~ 12	1416	354	1	1	1	1	
Het Nieuws van den Dag (17159617)	Training	1923	1, 2, 3, 4, 5, 6, 8, 10, 11, 12	1572	255	N.A.				
	Testing	1923	7,9	332	52	1	1	1	1	
		1924	4	138	26	1	1	1	1	
		1923	1 ~ 12	1904	367	1	1	1	1	
Vlaanderen (17164874)	Training	1923	1, 2, 4, 5, 6, 7, 8, 9, 10, 12	344	43	N.A.				
	Testing	1923	3,11	64	8	1	1	1	1	
		1924	7	32	4	1	1	1	1	
		1923	1 ~ 12	408	51	1	1	1	1	
De Volksgazet (17165064)	Training	1923	1, 2, 3, 4, 5, 6, 7, 9, 11, 12	1246	260	N.A.				1923-08-06-Pages1
	Testing	1923	8,10	258	54	0.98	0.98	0.99	0.98	
		1924	6	134	25	1	1	1	1	
		1923	1 ~ 12	1504	314	1	0.997	0.998	0.998	
De Standaard (17172097)	Training	1923	1, 2, 3, 4, 7, 8, 9, 10, 11, 12	1552	298	N.A.				1923-05-03-Pages3 1923-07-29-Pages1 1923-08-11-Pages1
	Testing	1923	5,6	318	59	0.98	1	1	0.99	
		1924	5	168	30	1	1	1	1	
		1923	1 ~ 12	1870	357	0.997	0.994	0.996	0.996	
Independence Belge (17229173)	Training	1923	1, 2, 3, 4, 5, 6, 7, 8, 10, 12	1564	296	N.A.				1923-05-30-Pages5
	Testing	1923	9,11	314	57	1	1	1	1	
		1924	1	162	30	1	1	1	1	
		1923	1 ~ 12	1878	353	0.997	1	1	0.999	



BE-
KBR00_15981364_19230416_00_01_00_1_01_0001_11880311



BE-
KBR00_15981364_19230416_00_01_00_1_01_0003_11880313



BE-
KBR00_17172097_19230503_00_01_00_1_01_0001_20557224



BE-
KBR00_17172097_19230503_00_01_00_1_01_0003_20557226

PAGE INEXISTANTE
NICHTVORHANDENE SEITE
ONBESTAANDE BLADZIJDE

BE-
KBR00_17172097_19230503_00_01_00_1_01_0001_20557224

Towards Full-scale Training

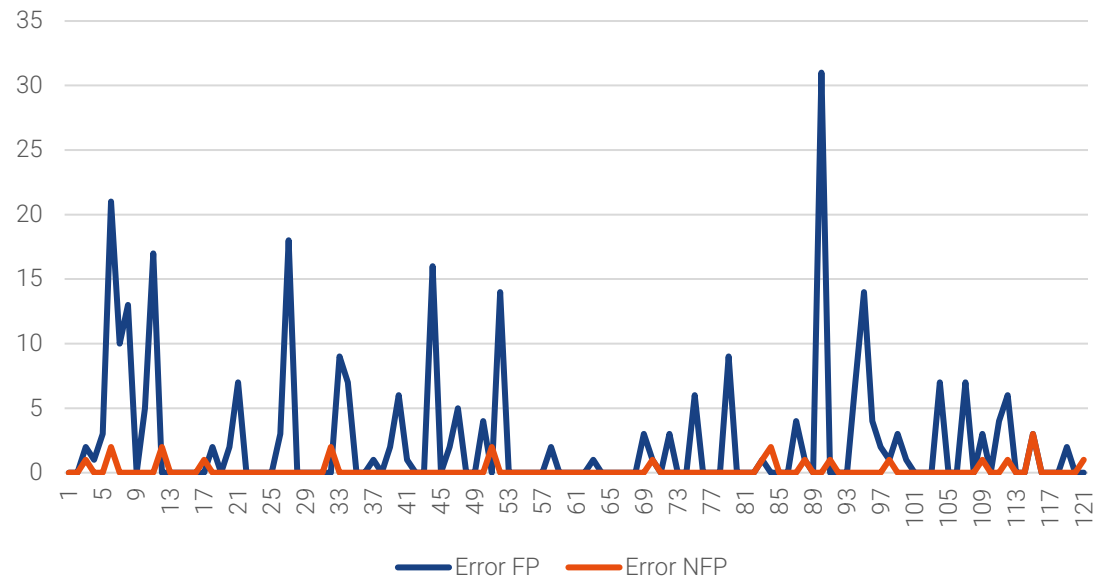
Training Configuration:

- ResNet-50, parameter size ~25.6M, basic CE loss, without hyperparameters optimization
- Prioritize on computation efficiency, input images are scaled to 224x224
- 121 different newspaper titles over the period of one year (80% training and 20% evaluation)
- ~130K pages, ~30K front pages

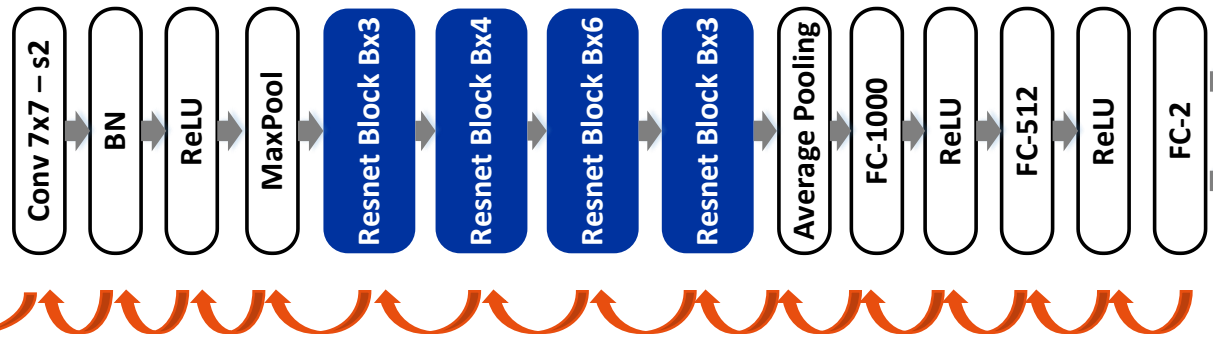
Performance (@50 training epochs):

- Precision 0.98, Recall 0.94, bAccuracy 0.97, F1 0.96
- ~65 newspaper titles: 0 mistakes
- ~10 newspaper titles: more than 5 mistakes
- ~ 8 newspaper titles: more than 10 mistakes

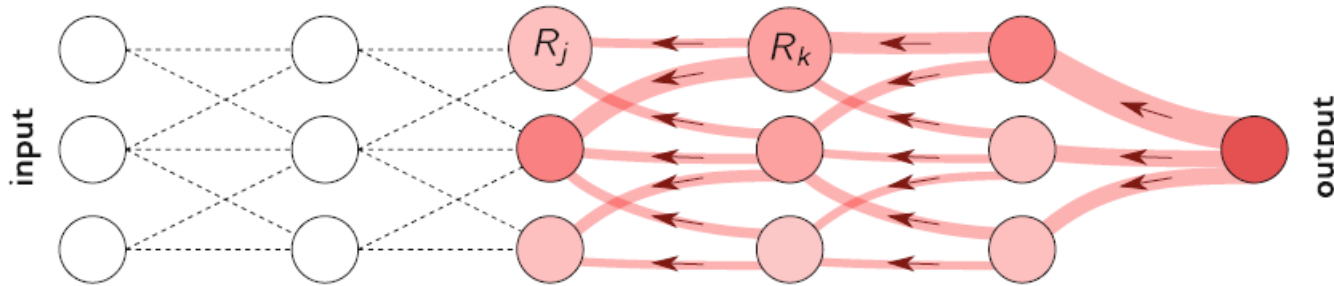
Distribution of Absolute Number of Errors



Deep learning with increased transparency: attribution + data analysis



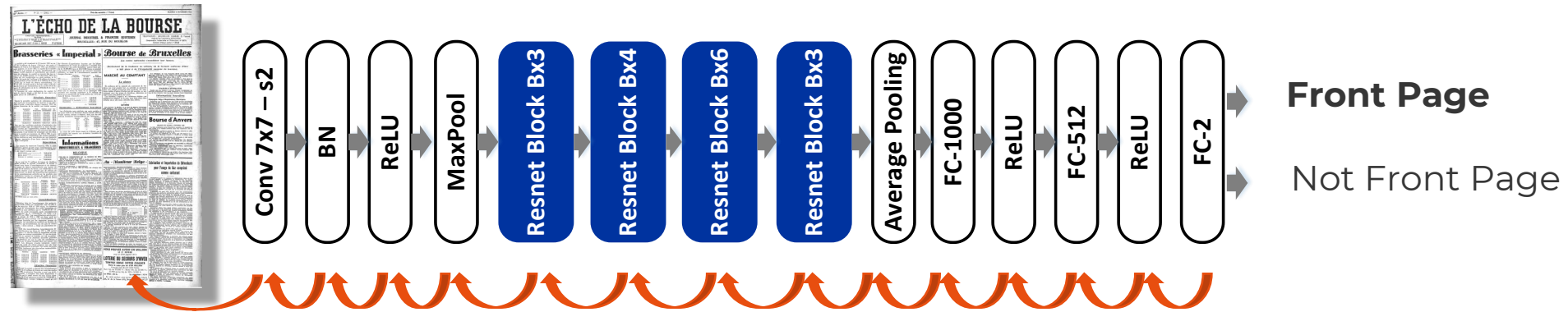
Front Page
Not Front Page



$$R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} R_k \rightarrow \sum_j R_j = \sum_k R_k$$

$$R_j = \sum_k \frac{a_j \rho(w_{jk})}{\epsilon + \sum_j a_j \rho(w_{jk})} R_k$$

Deep learning with increased transparency: attribution + data analysis

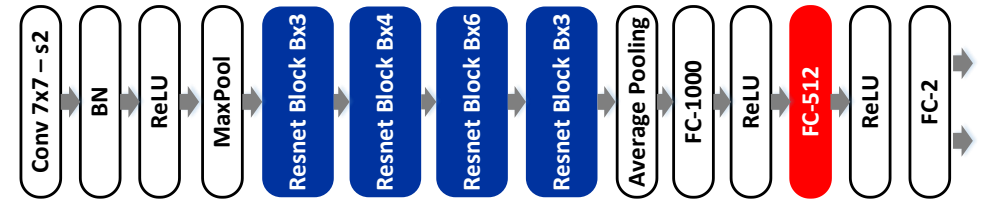
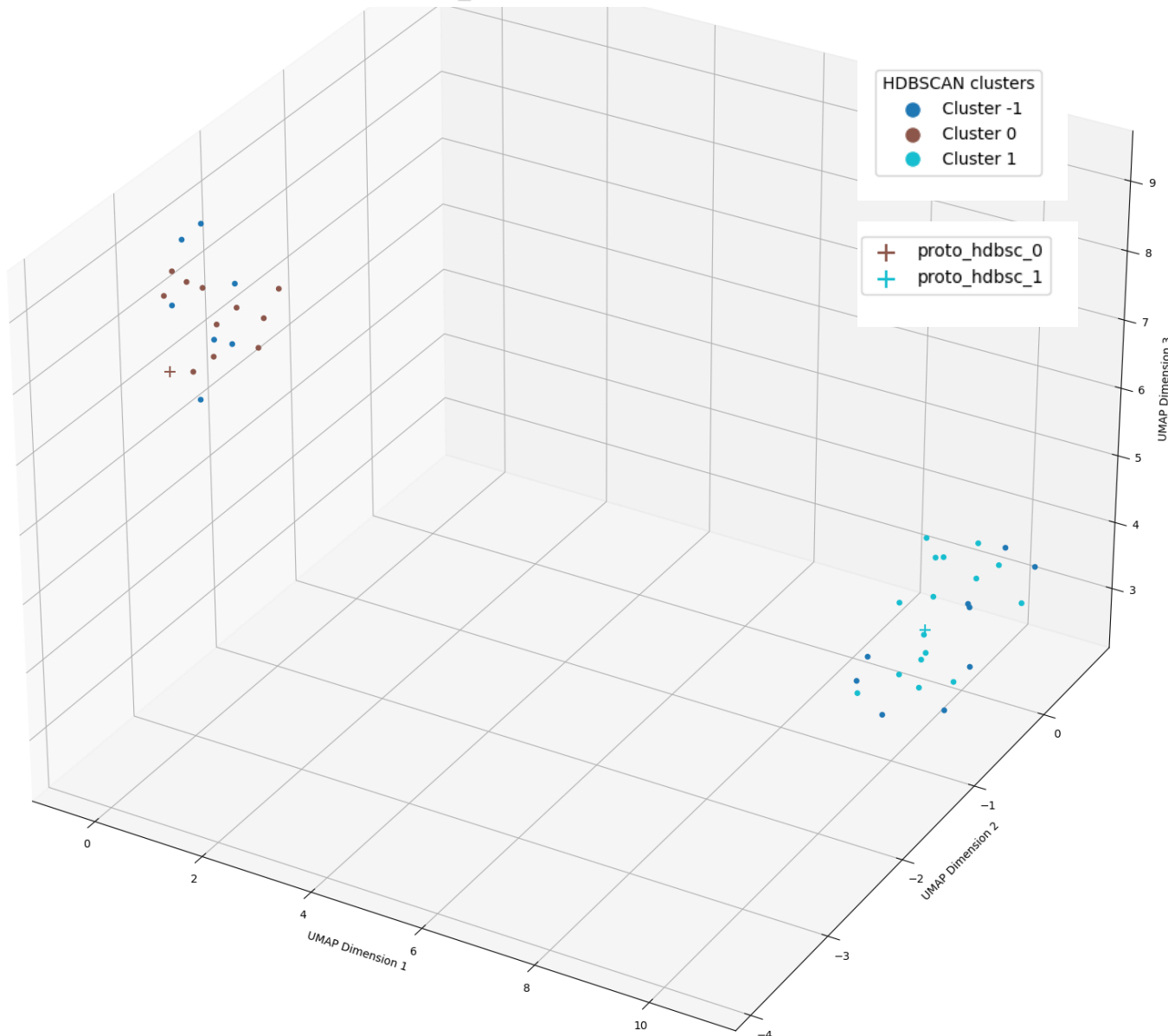


For each layer l , we could compute the relevance for all the neurons of the layer, resulting a relevance vector $r^{c \times h \times w}$, where channel c and spatial dimension of h, w are determined by the structure of the layer. For example, *layer 4.2 conv3* will yield a relevance vector of $r^{2048 \times 7 \times 7}$

- Review visual concepts that the neural network have learned
- Identify prototypes that are learned by the neural network
- Build an auxiliary data analysis pipeline to enhance the operation of the AI agent

Deep learning with increased transparency: attribution + data analysis

HDBSCAN clusters on channel_rel for 17145890 - fc.2 (43 samples)

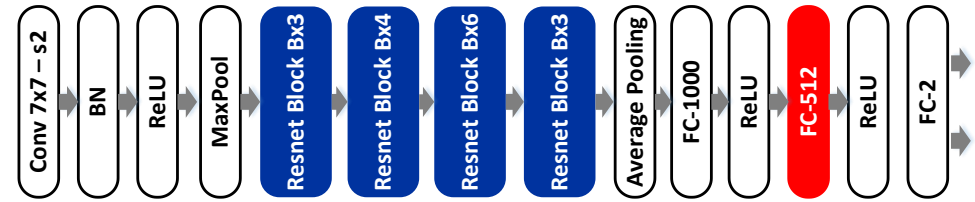


Analysis of all samples of a same newspaper title based on relevance vector r^{512} yielded by the second-last FC layer

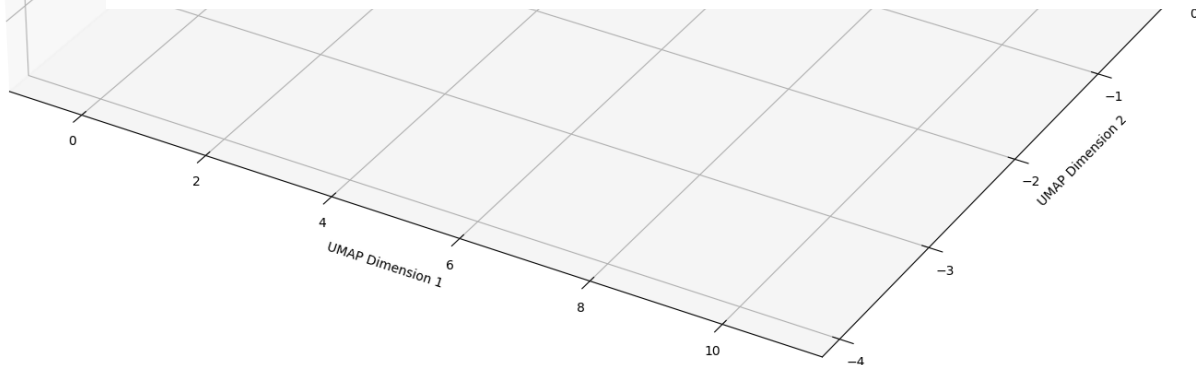
- Hierarchical density-based spatial clustering reveals two distinct clusters of samples
- We further compute the Euclidean centers of the two clusters
- Based on the computed centers, we identify representative samples (i.e., samples that have shortest distances to the centers)
- Visualization of the sample clusters with uniform manifold approximation and projection (UMAP)
- Review of the representative samples explains the clustering

Deep learning with increased transparency: attribution + data analysis

HDBSCAN clusters on channel_rel for 17145890 - fc.2 (43 samples)

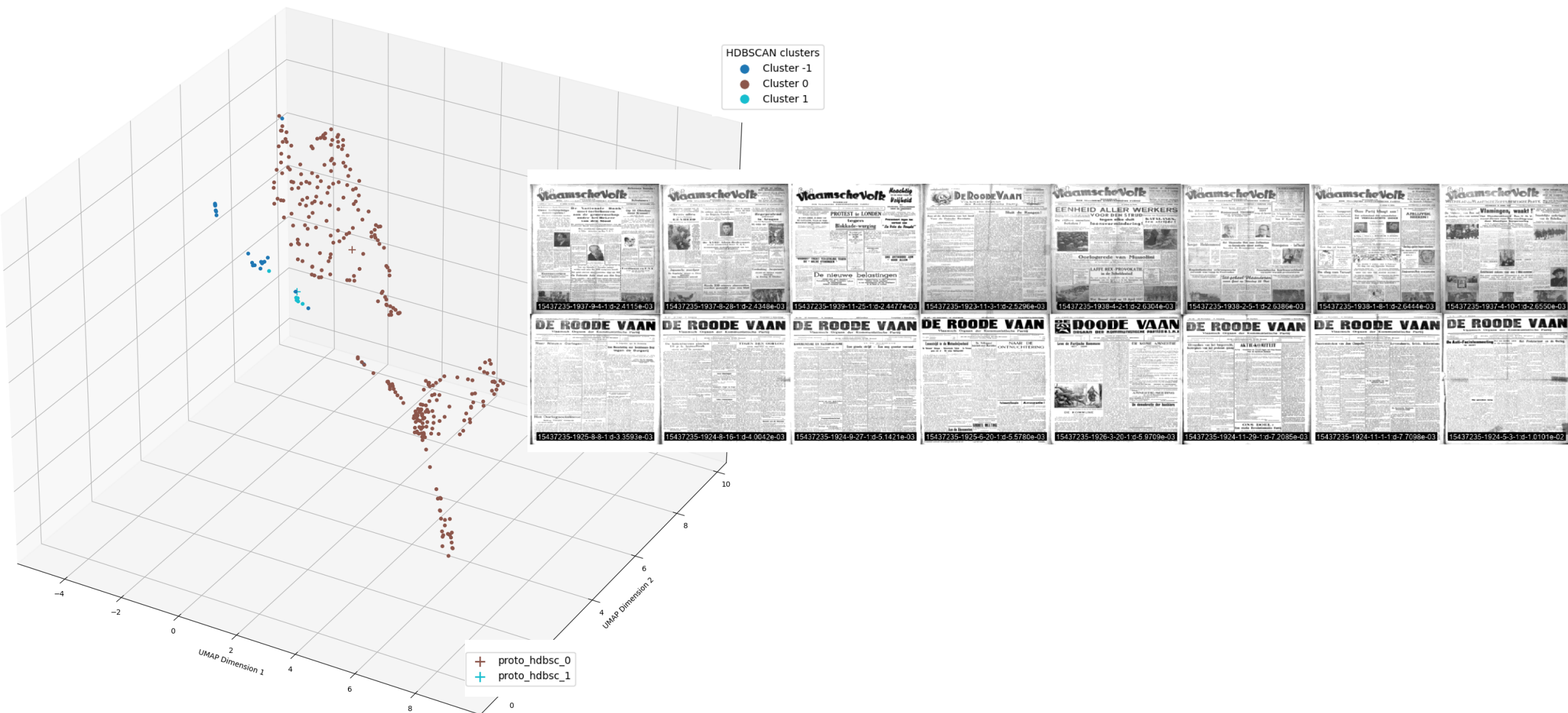


HDBSCAN clusters
● Cluster -1
● Cluster 0
● Cluster 1



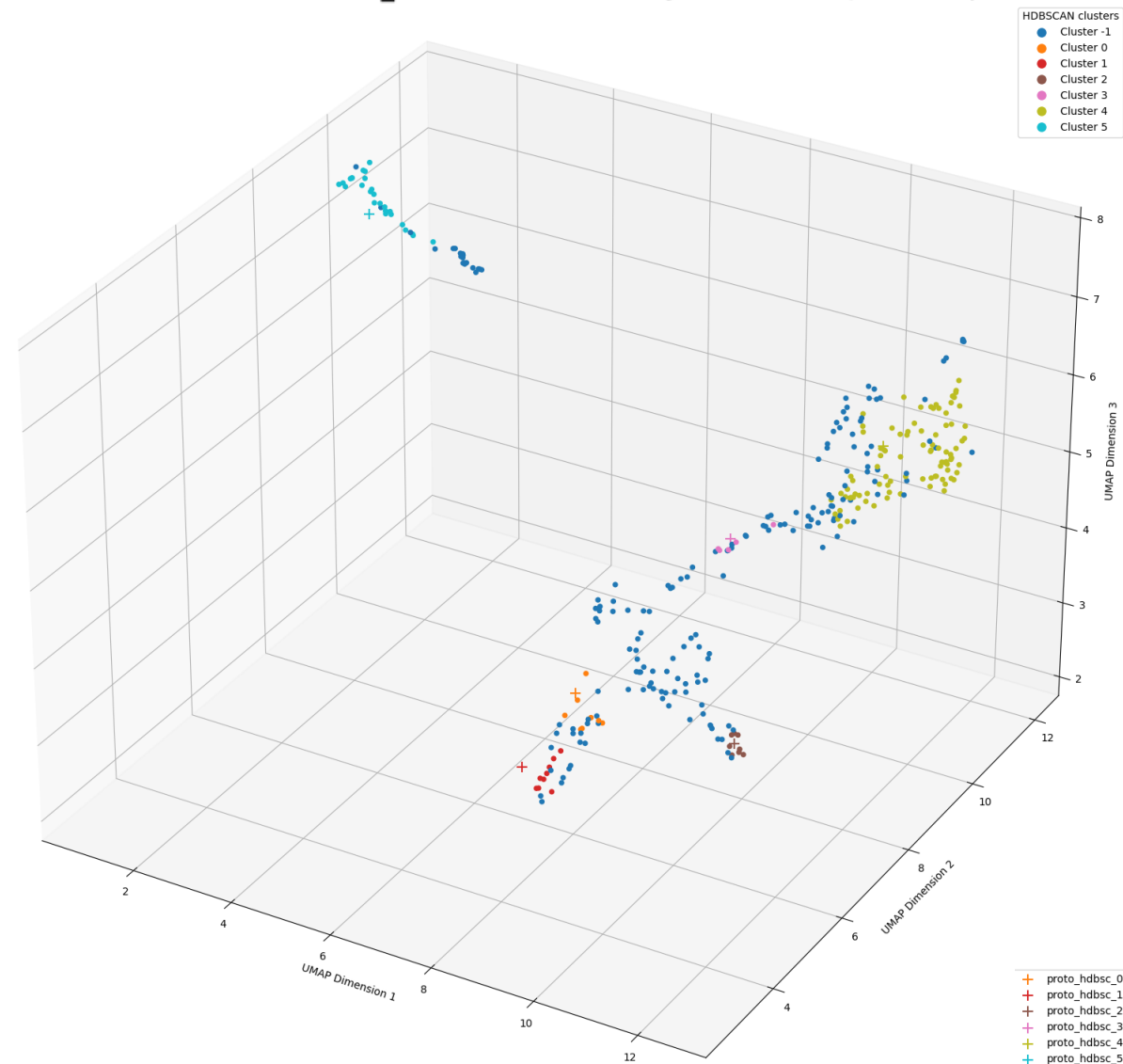
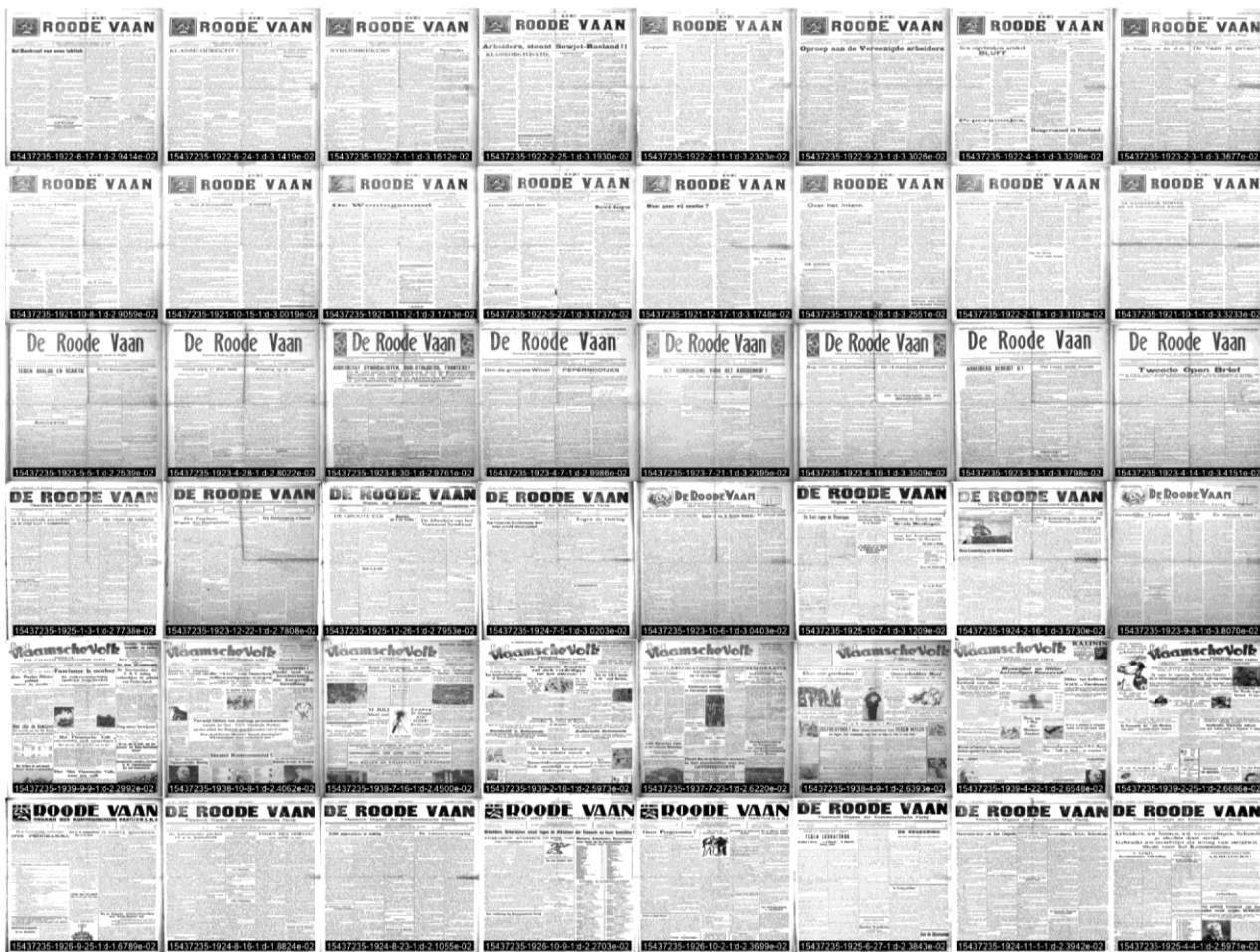
Deep learning with increased transparency: attribution + data analysis

HDBSCAN clusters on channel_rel for 15437235 - fc.2 (332 samples)

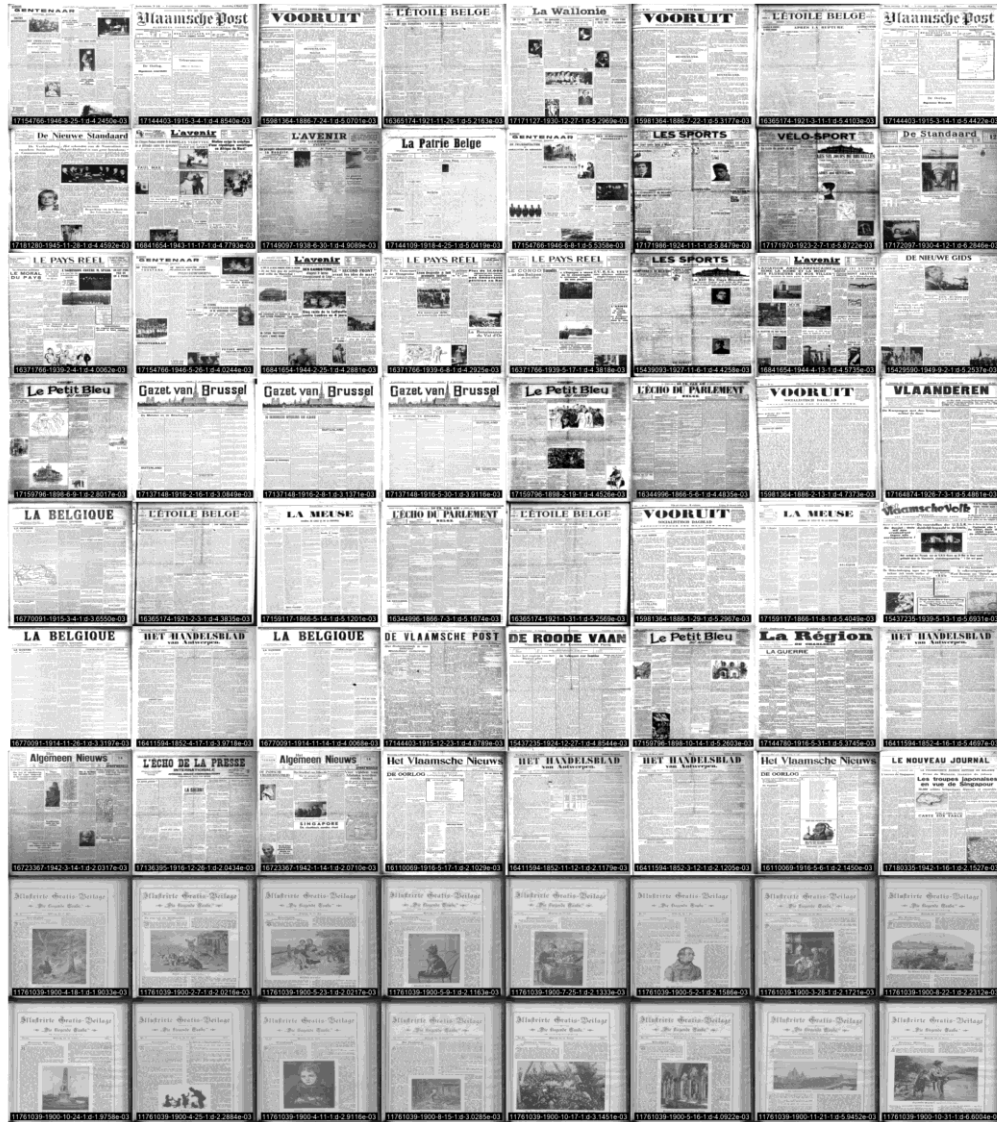


Deep learning with increased transparency: attribution + data analysis

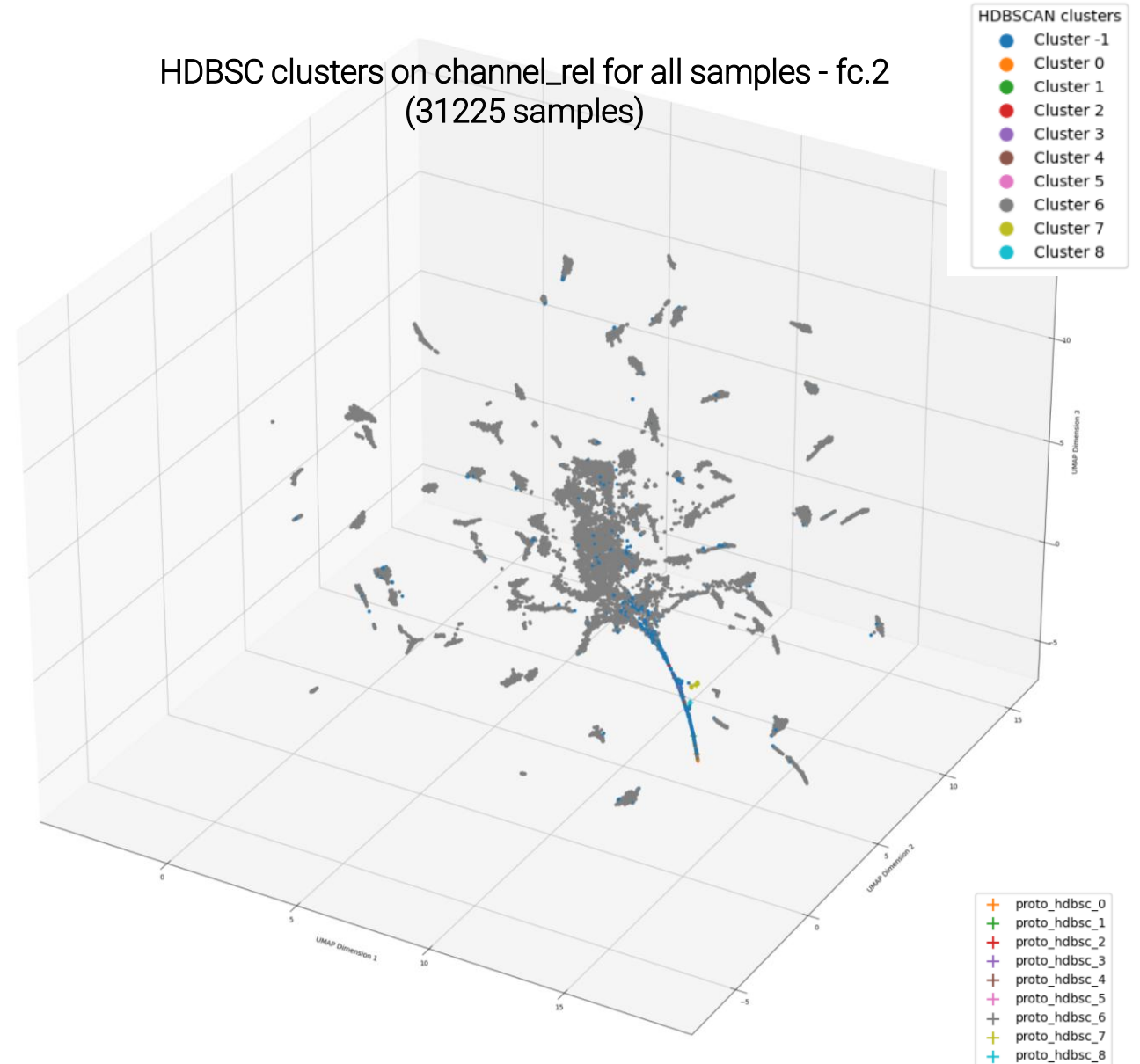
HDBSCAN clusters on channel_rel for 15437235 - layer4.2.conv3 (332 samples)



Deep learning with increased transparency: attribution + data analysis



HDBSCAN clusters on channel_rel for all samples - fc.2
(31225 samples)



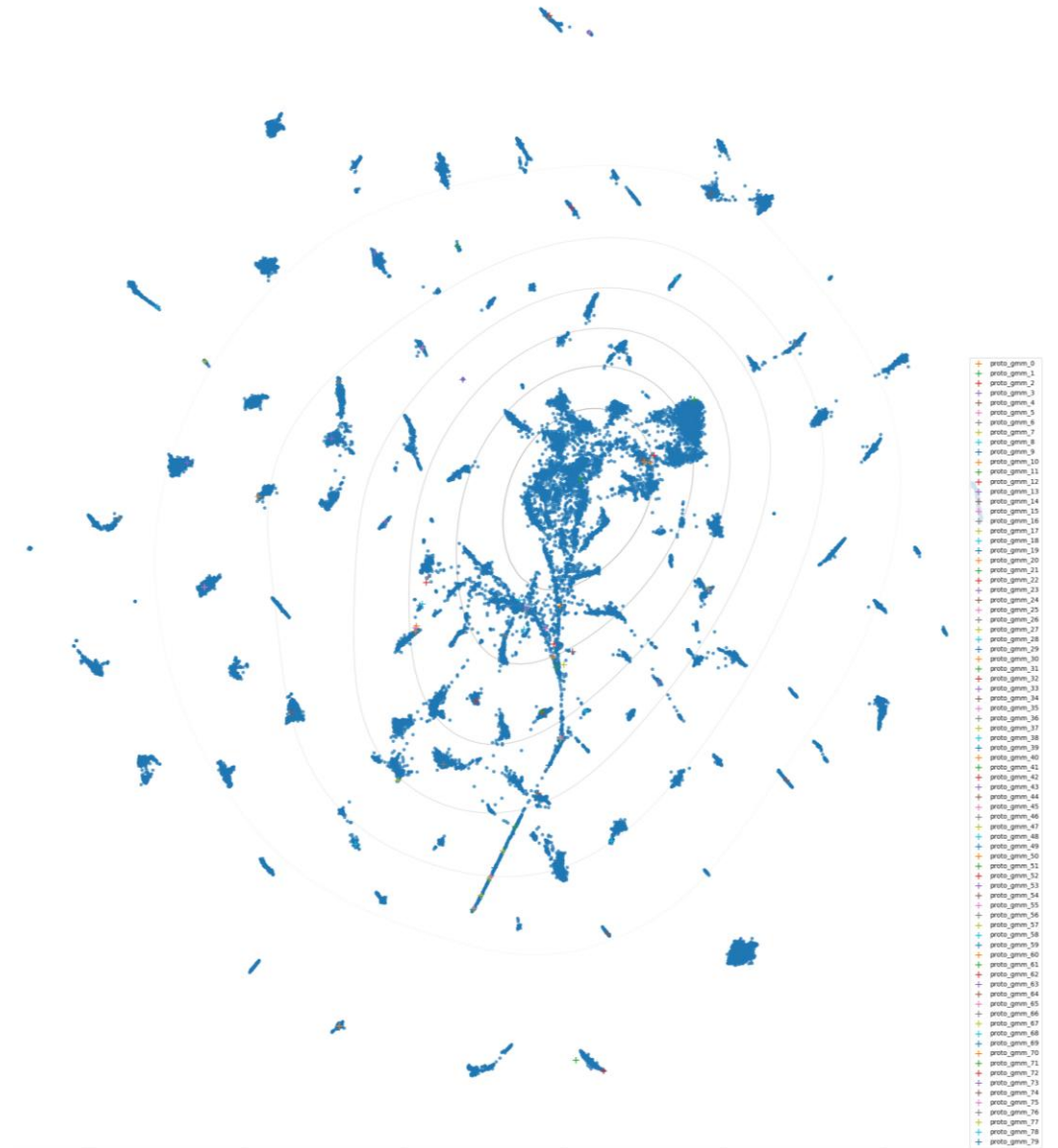
Deep learning with increased transparency: attribution + data analysis

More complicated modelling with Gaussian Mixture (GM) (80 different kernels)

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

$$0 < \pi_k < 1, \sum_{k=1}^K \pi_k = 1$$



Deep learning with increased transparency: attribution + data analysis

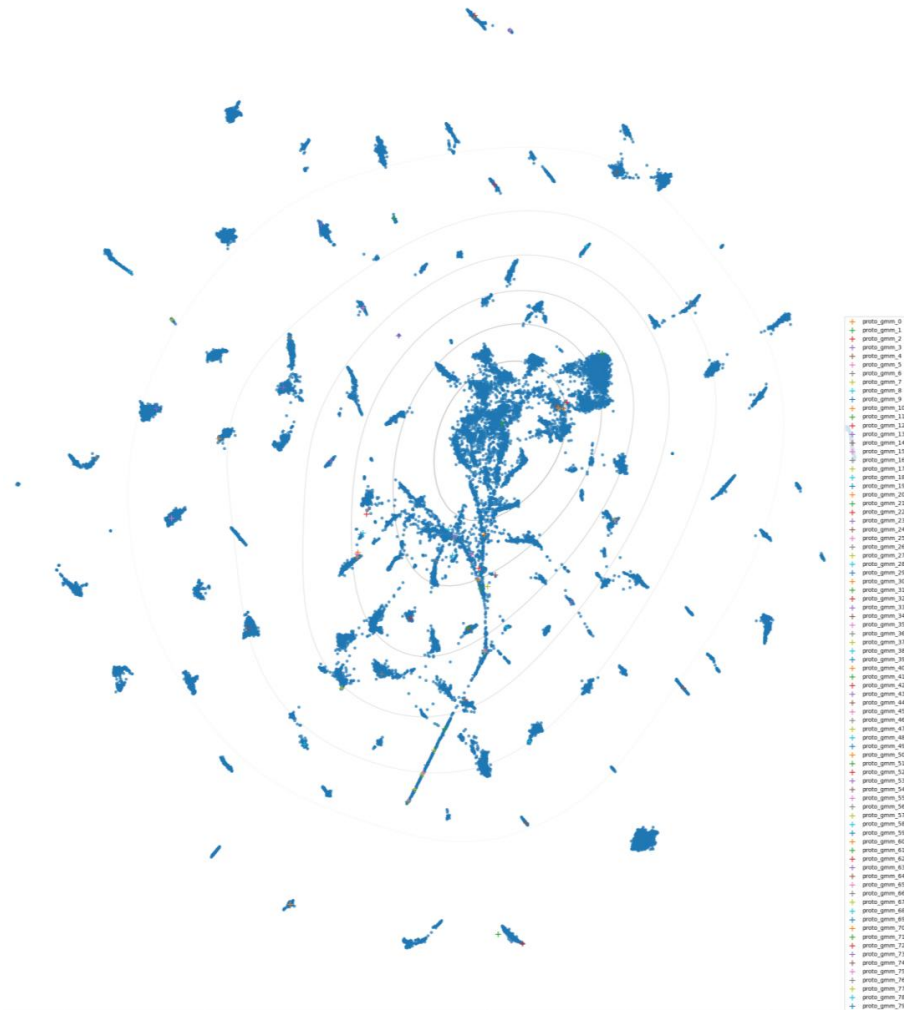
Representative samples for GM prototypes



Deep learning with increased transparency: attribution + data analysis

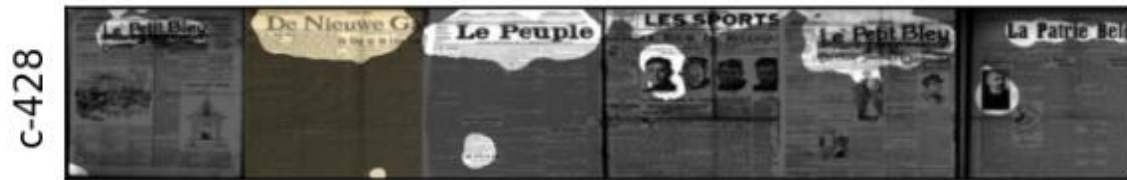
From the prototypes, we can also identify most relevant **visual concepts (VCs)**, these VCs help to address questions such as:

- Based on which regions of the newspapers is the neural network making decisions on its FP vs NFP classification?



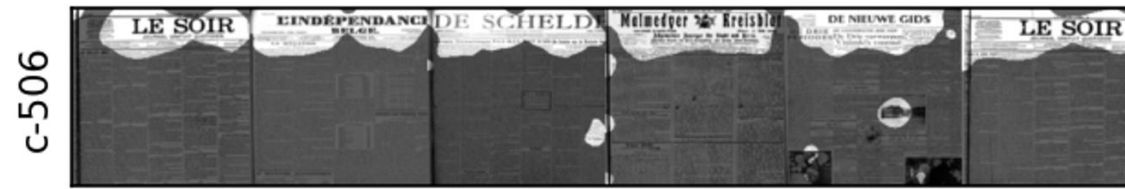
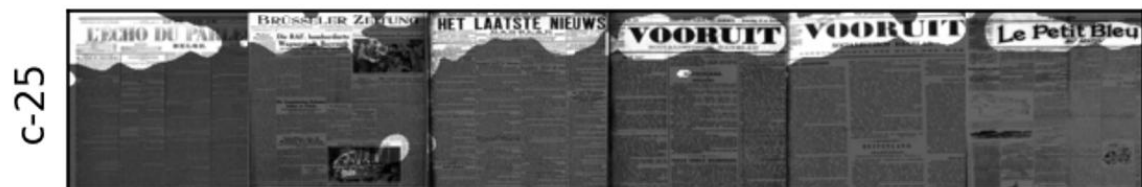
Deep learning with increased transparency: attribution + data analysis

Visualization of VCs (extracted from HDBSC prototypes extracted on layer fc.2) learned by the neural network for recognizing front pages

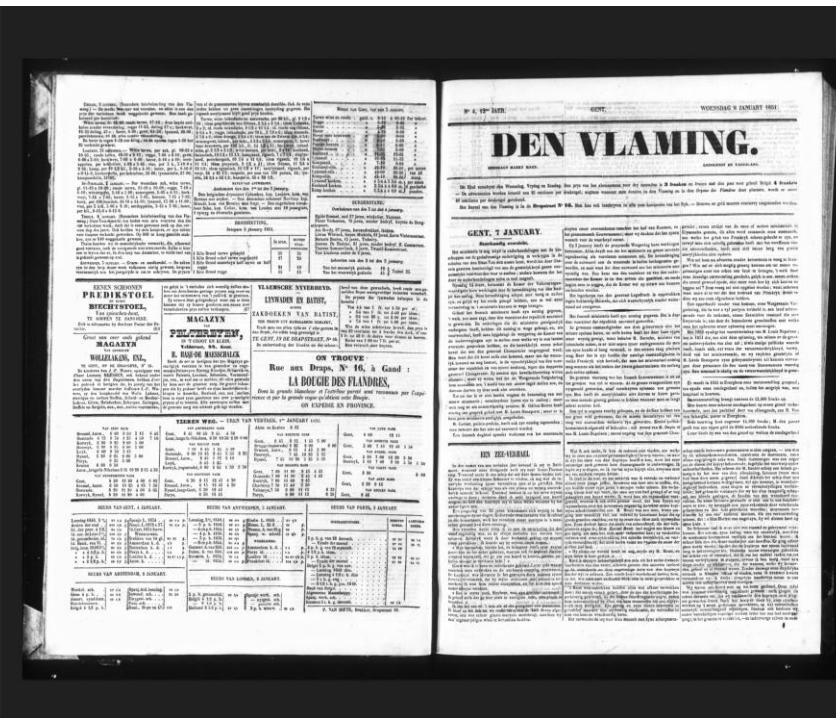
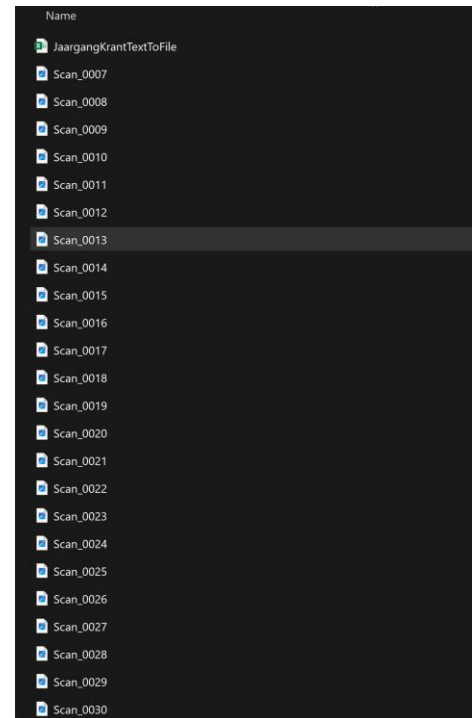
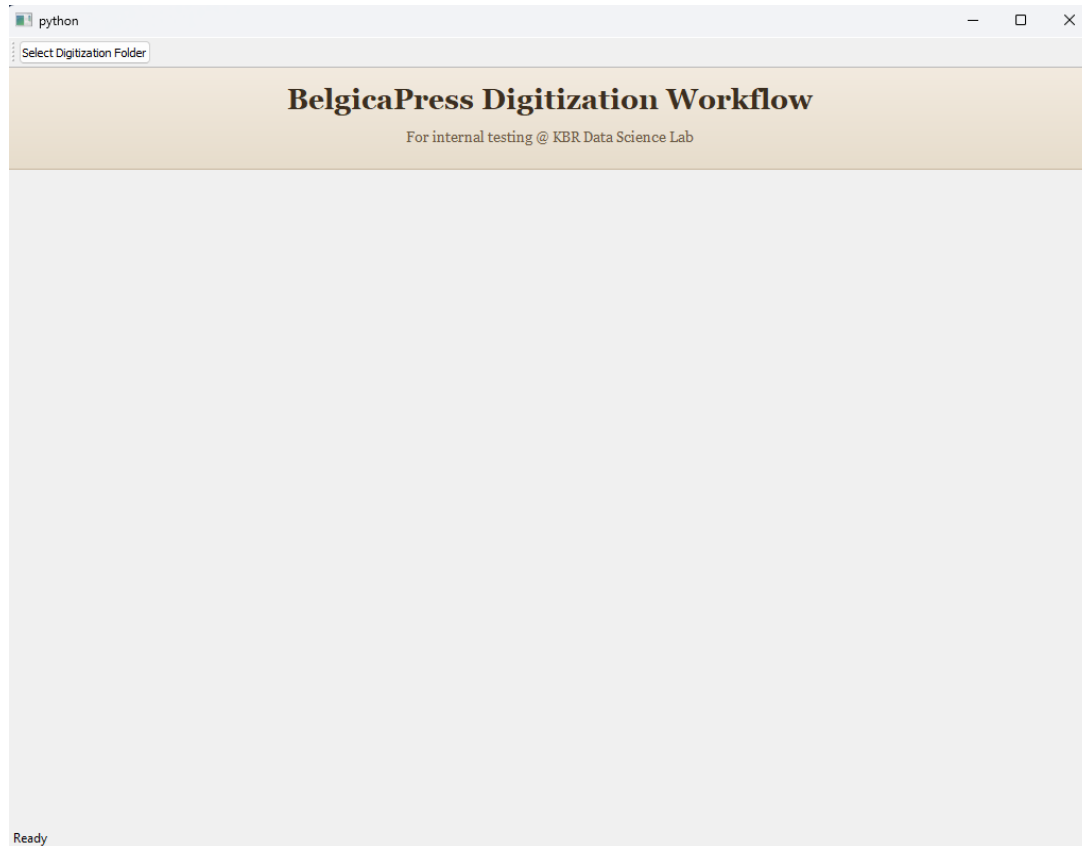


Deep learning with increased transparency: attribution + data analysis

Visualization of VCs (extracted from GMM prototypes on layer fc.2) learned by the neural network for recognizing front pages



From model development to workflow automation



From model development to workflow automation

The screenshot shows a web application window titled "python" with a sub-header "Select Digitization Folder". The main heading is "BelgicaPress Digitization Workflow" with the subtitle "For internal testing @ KBR Data Science Lab". The status is "Preparing dataset..." with the note "Sandbox creation and batched inference are running." Below this is a progress bar with three steps: "Create Sandbox" (completed), "Detect Front Pages" (in progress), and "Results Ready" (pending). At the bottom left, a status bar shows "Detecting front pages (batched)..." with a green progress indicator. At the bottom right, there is a button labeled "Results Ready — Review Pages".

The screenshot shows the same web application window. The status is now "Preparation complete." with the instruction "Click the button to start reviewing front pages." The progress bar shows "Create Sandbox" (completed), "Detect Front Pages" (completed), and "Results Ready" (pending). At the bottom left, a status bar shows "Collecting front-page records for review...". At the bottom right, the "Results Ready — Review Pages" button is still present.

From model development to workflow automation

python
Select Digitization Folder

BelgicaPress Digitization Workflow

For internal testing © KBR Data Science Lab

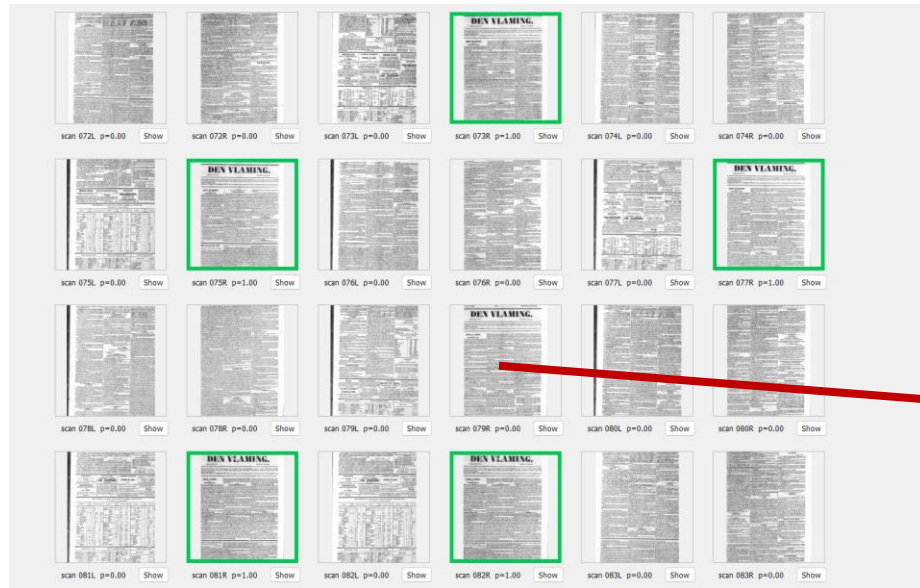
review detected front pages Inspector — scan 008 R p=1.00

Clear All Use Model Picks Confirm Selections

Review and confirm front pages.

Detailed view of a newspaper page from DEN VLAMING, dated VRYDAG 10 JANUARY 1851. The page features the title "DEN VLAMING." and various columns of text, including a section titled "Staatkundig overzicht." and a date "GENT, 9 JANUARY."

From model development to workflow automation



From model development to workflow automation

BelgicaPress Digitization Workflow

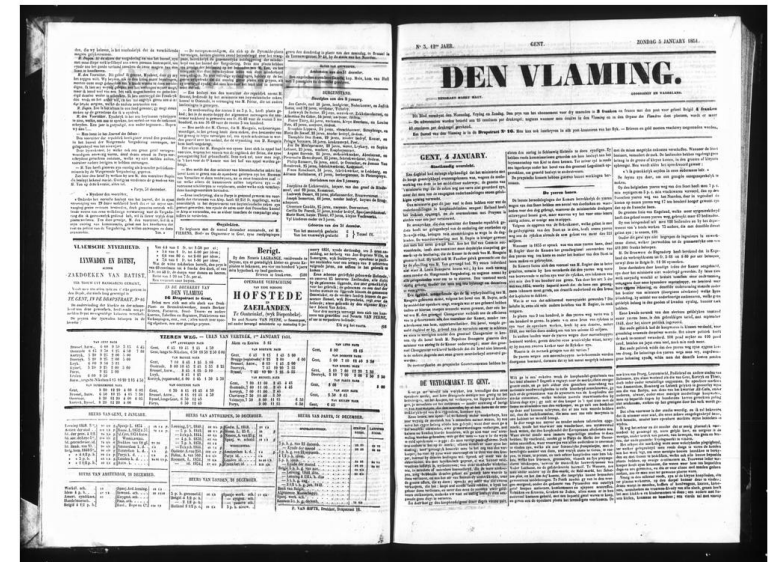
For internal testing @ KBR Data Science Lab

Proposed Edition Assignments

	Edition	Pages	Adjust
1	BE-KBR00_B-17146005_1851-01-01	 	Adjust
2	BE-KBR00_B-17146005_1851-01-02	  	Adjust
3	BE-KBR00_B-17146005_1851-01-03	 	Adjust
4	BE-KBR00_B-17146005_1851-01-04	  	Adjust
5	BE-KBR00_B-17146005_1851-01-05	  	Adjust
6	BE-KBR00_B-17146005_1851-01-06	  	Adjust
7	BE-KBR00_B-17146005_1851-01-07	  	Adjust
8	BE-KBR00_B-17146005_1851-01-08	  	Adjust

1 warning(s) detected

Inspector



Confirm Execute

From model development to workflow automation

Proposed Edition Assignments

	Edition	Pages	Adjust
40	BE-KBR00_B-17146005_1851-02-09		Adjust
41	BE-KBR00_B-17146005_1851-02-10		Adjust
42	BE-KBR00_B-17146005_1851-02-11		Adjust
43	BE-KBR00_B-17146005_1851-02-12		Adjust
44	BE-KBR00_B-17146005_1851-02-13		Adjust
45	BE-KBR00_B-17146005_1851-02-14		Adjust
46	BE-KBR00_B-17146005_1851-02-15		Adjust
47	BE-KBR00_B-17146005_1851-02-16		Adjust

Adjust Edition Assignments

Adjust membership for: Edition 44 — BE-KBR00_B-17146005_1851-02-13

Tip: checked = include this scan in the current edition. Others remain unchanged.

Edition 43: BE-KBR00_B-17146005_1851-02-12

2 scan(s)

[CURRENT] BE-KBR00_B-17146005_1851-02-13

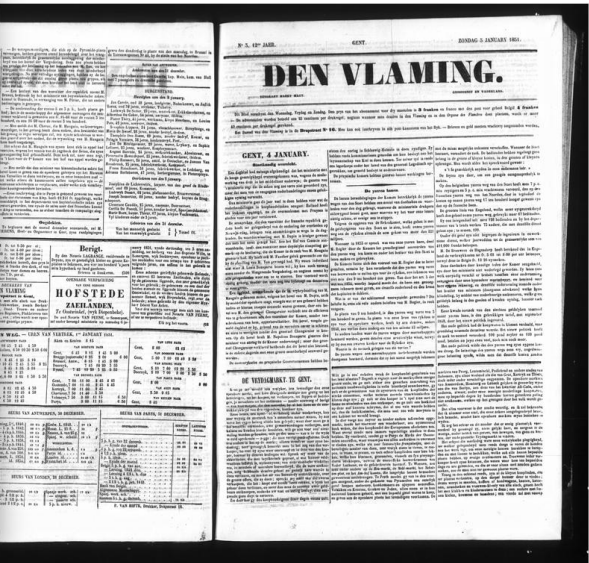
3 scan(s)

Edition 45: BE-KBR00_B-17146005_1851-02-14

3 scan(s)

Cancel
Apply

Inspector



▲ 1 warning(s) detected