



LIRMM

Panorama des outils de visualisation pour l'explicabilité en apprentissage profond pour le traitement automatique de la langue

Présenté par Alexis Delaforge

Alexis Delaforge^{1,3}

Jérôme Azé¹

Sandra Bringay^{1,2}

Arnaud Sallaberry^{1,2}

Maximilien Servajean^{1,2}

¹LIRMM, Université de Montpellier

²AMIS, Université Paul-Valéry

³Zortify Labs, Zortify



SIRIC
MONTPELLIER CANCER
Site de Recherche Intégrée sur le Cancer



INTERPRETABILITÉ & EXPLICABILITÉ

Interprétabilité = Transparence + Explicabilité

Zachary C. Lipton. The mythos of model interpretability : In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3) :31-57, 2018.



INTERPRETABILITÉ & EXPLICABILITÉ

Interprétabilité = Transparence + Explicabilité

Transparence : facilité avec laquelle un humain peut comprendre et reproduire le fonctionnement d'un modèle, indépendamment d'une prédiction.

Zachary C. Lipton. The mythos of model interpretability : In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3) :31-57, 2018.



INTERPRETABILITÉ & EXPLICABILITÉ

Interprétabilité = Transparence + Explicabilité

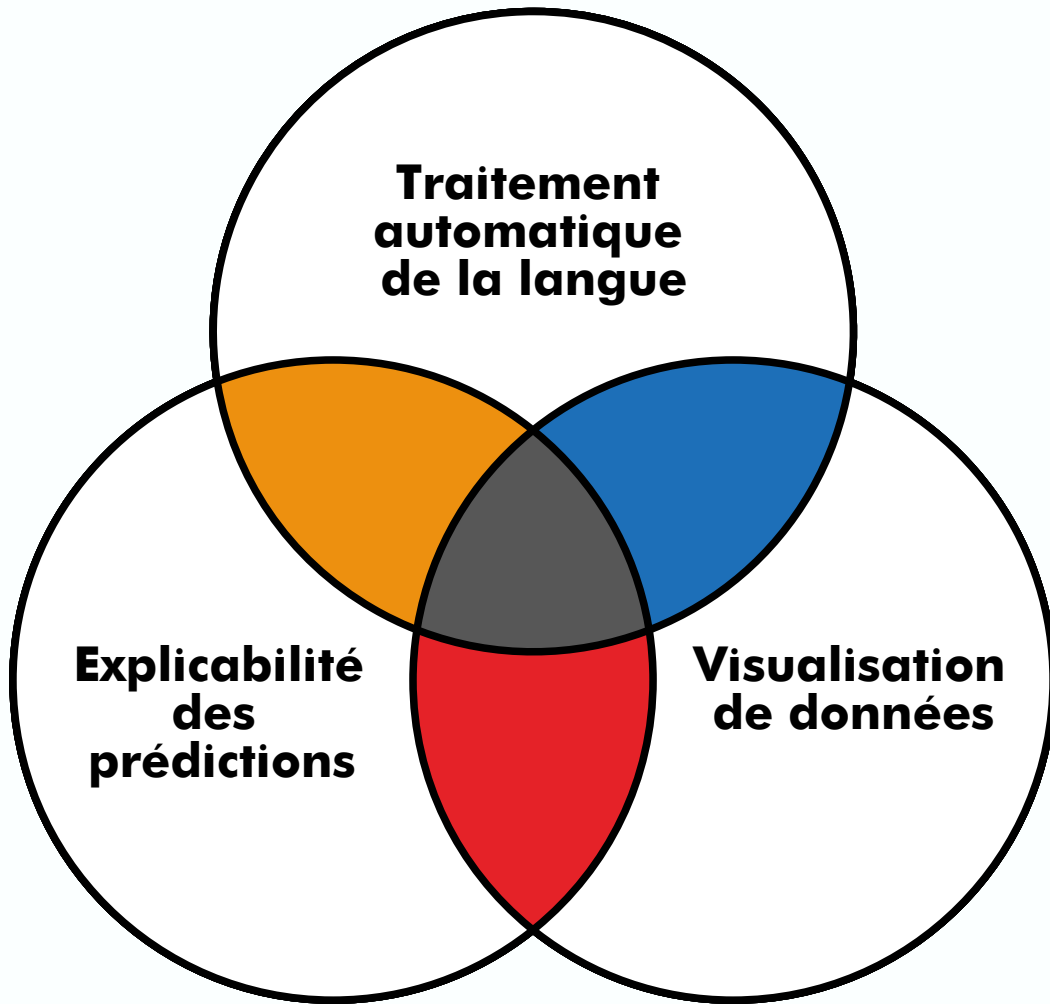
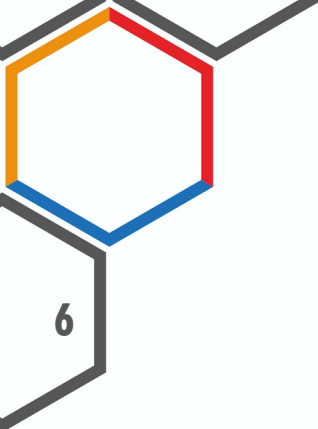
L'explication d'une prédiction, est faite lorsque des indicateurs, issus ou non du fonctionnement d'un modèle, sont utilisés pour expliquer sa décision.



INTERPRETABILITÉ & EXPLICABILITÉ

Pas de consensus :

- Zachary C. Lipton. The mythos of model interpretability : In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3) :31–57, 2018.
- Valérie Beaudouin, Isabelle Bloch, David Bounie, Stéphan Cléménçon, Florence d'Alché Buc, James Eagan, Winston Maxwell, Pavlo Mozharovskyi, and Jayneel Parekh. Flexible and context-specific ai explainability : a multidisciplinary approach. SSRN, 2020.
- ...



**Traitement
automatique
de la langue**

**Explicabilité
des
prédictions**

**Visualisation
de données**

Alharbi, Mohammad, and Robert S. Laramée. 2019. SoS TextVis: An Extended Survey of Surveys on Text Visualization. Computers 8, no. 1: 17. <https://doi.org/10.3390/computers8010017>

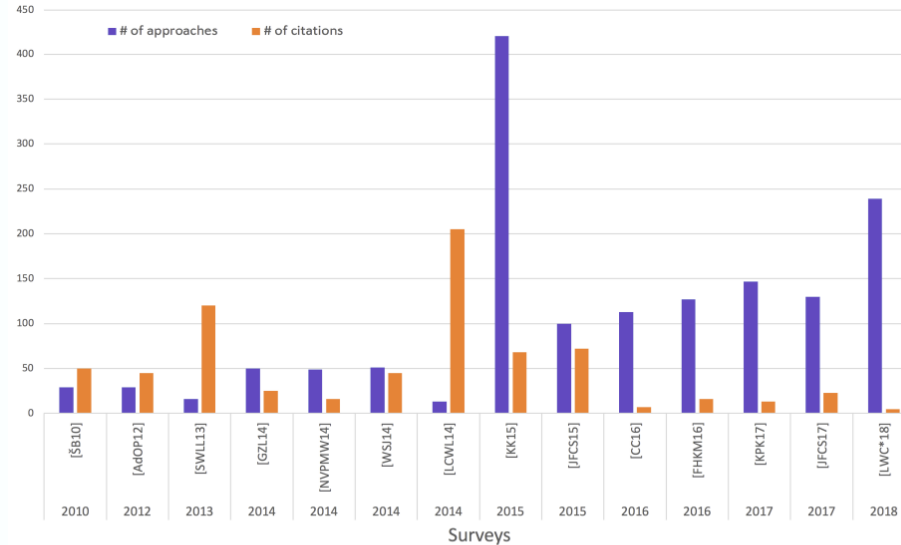
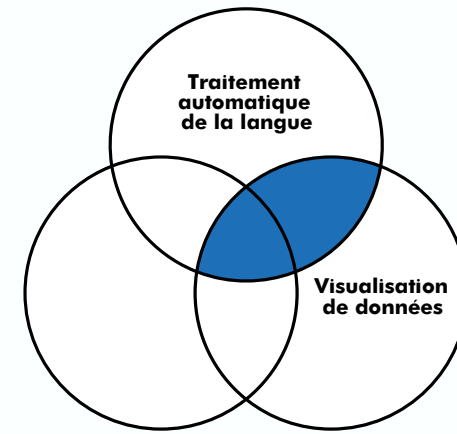
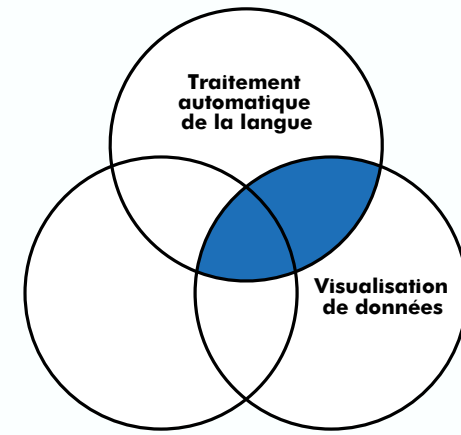


Figure 1. The text visualization surveys from 2010 to 2018. Blue bars indicate the number of methods reviewed in each survey. Orange bars show the number of citations each survey attracts. In term of the number of surveys, 2014 dominates with four surveys. However, with the respect to the number of techniques, surveys from 2015 review 480 methods collectively.

TAL & VIS

9

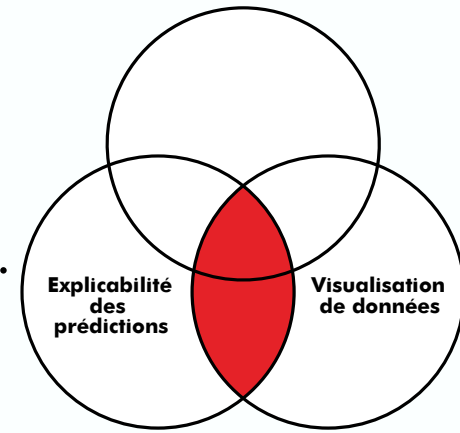
K. Kucher and A. Kerren, "Text visualization techniques: Taxonomy, visual survey, and community insights," 2015 IEEE Pacific Visualization Symposium (PacificVis), Hangzhou, China, 2015, pp. 117-121, doi: 10.1109/PACIFICVIS.2015.7156366.



<https://textvis.lnu.se/> jusqu'à 2019



Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. Visual analytics in deep learning : An interrogative survey for the next frontiers. IEEE TVCG, 25(8) :2674–2693, 2019.



Visual Analytics in Deep Learning | Interrogative Survey Overview

\$4 WHY

Why would one want to use visualization in deep learning?

- Interpretability & Explainability
- Debugging & Improving Models
- Comparing & Selecting Models
- Teaching Deep Learning Concepts

\$6 WHAT

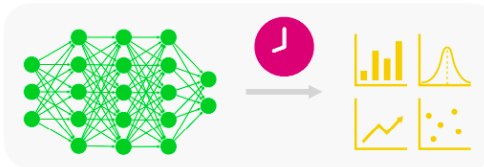
What data, features, and relationships in deep learning can be visualized?

- Computational Graph & Network Architecture
- Learned Model Parameters
- Individual Computational Units
- Neurons In High-dimensional Space
- Aggregated Information

\$8 WHEN

When in the deep learning process is visualization used?

- During Training
- After Training



\$5 WHO

Who would use and benefit from visualizing deep learning?

- Model Developers & Builders
- Model Users
- Non-experts

\$7 HOW

How can we visualize deep learning data, features, and relationships?

- Node-link Diagrams for Network Architecture
- Dimensionality Reduction & Scatter Plots
- Line Charts for Temporal Metrics
- Instance-based Analysis & Exploration
- Interactive Experimentation
- Algorithms for Attribution & Feature Visualization

\$9 WHERE

Where has deep learning visualization been used?

- Application Domains & Models
- A Vibrant Research Community

Fig. 1. A visual overview of our interrogative survey, and how each of the six questions, "Why, Who, What, How, When, and Where," relate to one another. Each question corresponds to one section of this survey, indicated by the numbered tag, near each question title. Each section lists its major subsections discussed in the survey.

XAI & VIS

Biagio La Rosa, Graziano Blasilli, Romain Bourqui, David Auber, Giuseppe Santucci, Roberto Capobianco, Enrico Bertini, Romain Giot, and Marco Angelini. State of the art of visual analytics for explainable deep learning. Computer Graphics Forum, 42(1) :319–355, 2023.

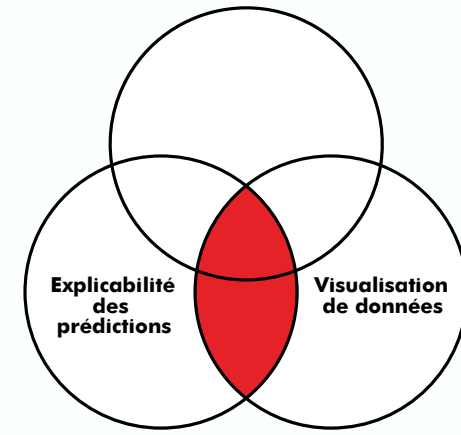


Table 1: List of the 67 Visual Analytics systems considered in this state-of-the-art report. The systems belong to five XDL categories: feature attribution (FA), Learned features (LF), explanation by examples (EE), counterfactuals examples (CE) and model behaviour (MB). Target users can be architects (A), trainers (T), and end users (E). Interactivity can be: passive (P), interactive input observations (I), and interactive model observations (M). Plus can be: training (TR) and testing (TE). Evaluation can be: quantitative user study (Q-US), user study with feedback (F-US), case study with feedback (F-CS), case study (CS) and usage scenarios (US). Furthermore, the table reports whether the authors have provided the source code of a system. Table 2 shows additional aspects of the considered systems.

#	System Name	Reference	Year	Category	FA	LF	EE	CE	MB	Application Domain	DL Model	Class	User	Interactivity	Plus	Evaluation					Code	
																Q-US	F-US	F-CS	CS	US		
01	ACE	[137]19	2019	FA						Medical	DL	Arch										
02	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
03	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
04	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
05	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
06	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
07	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
08	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
09	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
10	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
11	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
12	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
13	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
14	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
15	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
16	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
17	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
18	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
19	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
20	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
21	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
22	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
23	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
24	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
25	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
26	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
27	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
28	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
29	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
30	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
31	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
32	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
33	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
34	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
35	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
36	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
37	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
38	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
39	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
40	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
41	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
42	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
43	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
44	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
45	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
46	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
47	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
48	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
49	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
50	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
51	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
52	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
53	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
54	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
55	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
56	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
57	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
58	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
59	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
60	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
61	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
62	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
63	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
64	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
65	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
66	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
67	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
68	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
69	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
70	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
71	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
72	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
73	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
74	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
75	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
76	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
77	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
78	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
79	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
80	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
81	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
82	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
83	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
84	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
85	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
86	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
87	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
88	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
89	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
90	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
91	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
92	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
93	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
94	AdaptVis	[82]19	2019	FA						Medical	DL	Arch										
95	AdaptVis	[82]19	2019	FA						Medical	DL											

Angelos Chatzimparmpas, Rafael M. Martins, Ilir Jusufi, and Andreas Kerren. A survey of surveys on the use of visualization for interpreting machine learning models. Information Visualization, 19(3) :207–233, 2020.

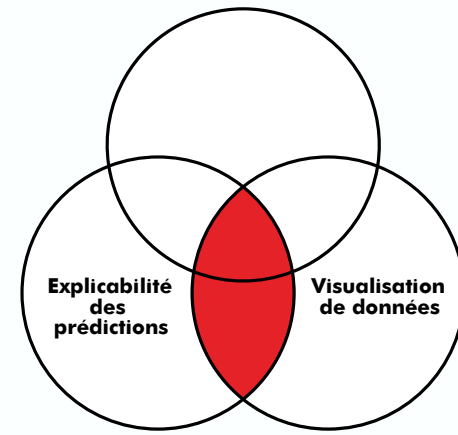


Table 6. Top eight terms for each of the 10 topics generated with latent Dirichlet allocation (LDA) from the collection of the papers of the survey papers.

Topic 1 <i>NNs for image applications</i>	Topic 2 <i>IML classifiers' training</i>	Topic 3 <i>DR and projections visualization</i>	Topic 4 <i>CNNs (for image applications)</i>	Topic 5 <i>Use of features in models' predictions (for regression)</i>
Image Network Layer Feature Input Model Neural Deep	Model User Learning Training Machine Classifier System Machine learning	User Visualization Dimension Point Analysis Visual Projection Uncertainty	Part Object CNN Image Node Pattern Layer Category	Feature Model Function Point Prediction Value Vector Regression
Topic 6 <i>Subspace visualization for clustering and classification</i>	Topic 7 <i>Users' feedback in ML</i>	Topic 8 <i>Clustering and algorithm use in feature subset selection</i>	Topic 9 <i>Model and clustering visualization for time-series data</i>	Topic 10 <i>NNs for text applications</i>
Cluster Visualization Class Clustering Subspace Model Analysis Classification	User Topic Feature Learning Document Graph System Participant	Feature Cluster Selection Clustering Algorithm Subset Search Learning	Model Cluster User Visualization Time Visual Analysis Clustering	Model Image Network Learning Word Neural State Training

NN: neural network; IML: interactive machine learning; DR: dimensionality reduction; CNN: convolutional neural networks; ML: machine learning.

The suggested topic titles are displayed in italics. Each topic is represented by one specific color.

Angelos Chatzimparmpas, Rafael M. Martins, Ilir Jusufi, and Andreas Kerren. A survey of surveys on the use of visualization for interpreting machine learning models. *Information Visualization*, 19(3) :207–233, 2020.

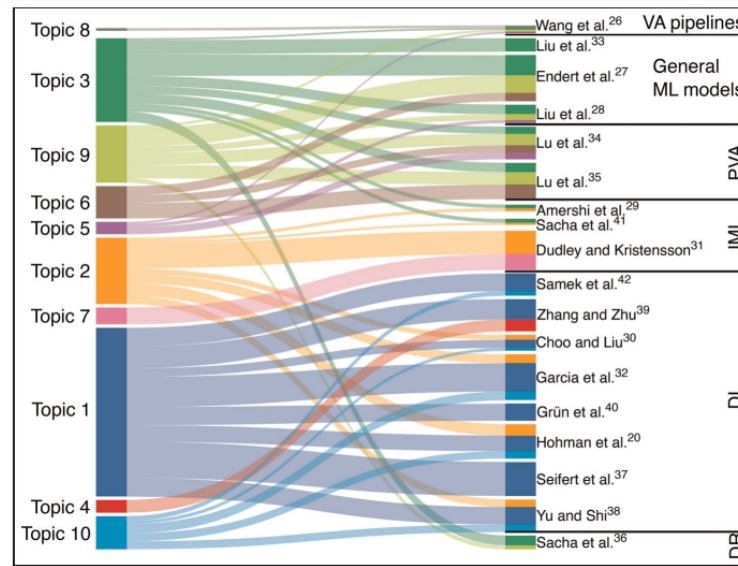
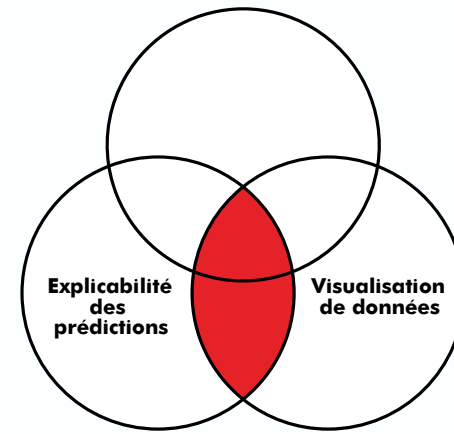
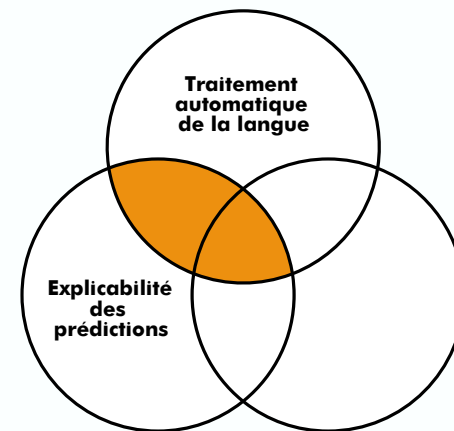


Figure 4. Explicit connections between the survey papers and the LDA topics extracted automatically from the individual papers under each survey. The width of the connections shows the weight of each topic in each survey paper. On the right side, the survey papers are categorized and sorted according to Table 5. On the left side, the topics are ordered to minimize the number of crossing lines.

Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc interpretability for neural nlp : A survey. ACM Computing Surveys, 55(8) :1-42, 2022.



		less information			more information		
		post-hoc			intrinsic		
		black-box	dataset	gradient	embeddings	white-box	model specific
lower abstraction	local explanation						
	input features	SHAP § A.2	LIME § 6.2, Anchors § A.3	Gradient § 6.1, IG § A.1			Attention
	adversarial examples	SEA ^M § B.1	HotFlip § 7.1				
	influential examples	Influence Functions ^H § 8.1		TracIn ^C § 8.3		Representer Pointers [†] § 8.2	Prototype Networks
	counter-factuals	Polyjuice ^{M,D} § C.1	MiCE ^M § 9.1				
natural language	CAGE ^{M,D} § 10.1						GEF ^D , NILE ^D
higher abstraction	class explanation						
	concepts						NIE ^D § 11.1
	global explanation						
	vocabulary				Project § 12.1, Rotate § 12.2		
ensemble	SP-LIME § 13.1						
linguistic information	Behavioral Probes ^D § 14.1				Structural Probes ^D § 14.2	Structural Probes ^D § 14.2	Auxiliary Task ^D
rules	SEAR ^M § 15.1	Compositional Explanations of Neurons [†] § D.1					

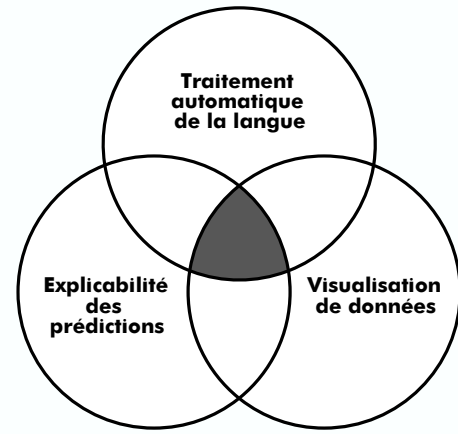
Rows describe how the explanation is communicated, while columns describe what information is used to produce the explanation. The order of both rows and columns indicates a level of abstraction and amount of information, respectively. However, this order is only approximate.

Furthermore, because this survey focuses on *post-hoc* methods, the *intrinsic* section of this table is incomplete and merely meant to provide a few comparative examples. The specific *intrinsic* methods shown are: *Attention* [9], *GEF* [69], *NILE* [63]. *Prototype Networks* and *Auxiliary Task* refer to types of models.

^C: Depends on checkpoints during training. ^D: Depends on supplementary dataset. ^H: Depends on second-order derivative. ^M: Depends on the supplementary model. [†]: Depends only on the dataset and white-box access.

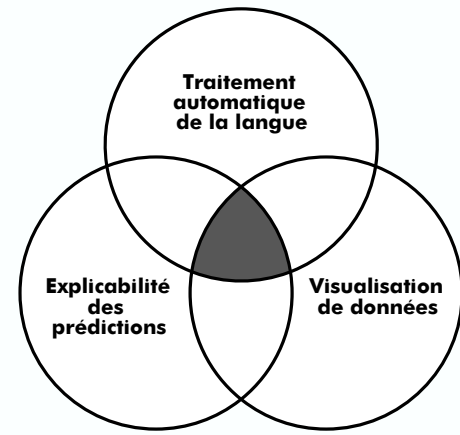
TAL, VIS & XAI

Il n'existe pas à ce jour de travaux s'intéressant aux techniques de visualisation utilisées pour expliquer les prédictions dans le domaine du traitement automatique de la langue.



TAL, VIS & XAI

Il n'existe pas à ce jour de travaux s'intéressant aux techniques de visualisation utilisées pour expliquer les prédictions dans le domaine du traitement automatique de la langue.



Types de réseaux :

- Perceptrons multicouches et réseaux convolutifs
- Réseaux de neurones récurrents
- Réseaux auto-attentionnels
- Méthodes agnostiques

PERCEPTRONS MULTICOUCHES ET RÉSEAUX CONVOLUTIFS

northrop grumman corp on monday said it received a 10 year \$ 408 million army contract to provide simulated battle command training support to army corps commanders the latest award in

northrop grumman corp on monday said it received a 10 year \$ 408 million army contract to provide simulated battle command training support to army corps commanders the latest award in

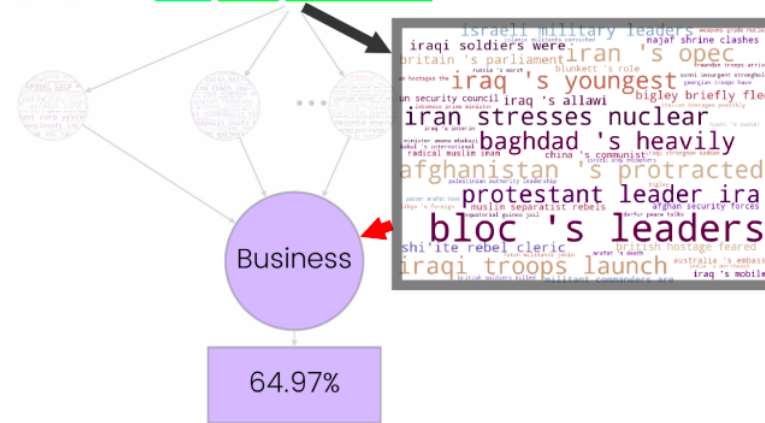
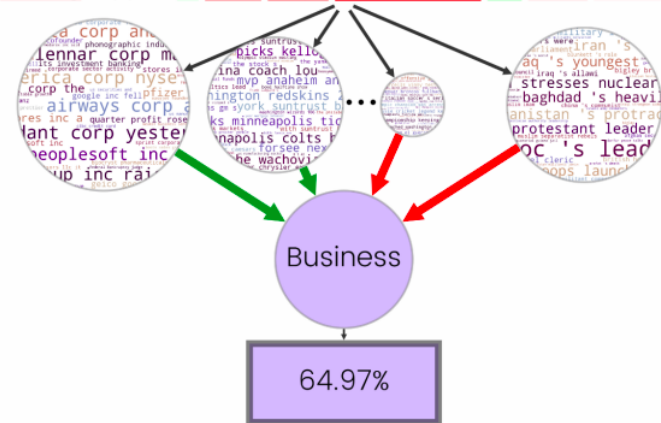


Figure 4: (Left) Graphical DAX for the (correct) output class “Business” (determined with 64.97% probability) by the CNN for text classification with AG-News, for the input article from AG-News at the top. (Right) The view of the DAX after clicking on the rightmost argument/word cloud. Arguments are visualised (by χ) as follows: *input words* with highlighting of a word given, on the left, by the sum of the (dialectical) strengths of all arguments representing that word and, on the right, by its corresponding argument’s dialectical strength; *convolutional filters/word clouds* with sizes the corresponding arguments’ strengths; and the most probable *class* and its probability. The green/red colours indicate, resp., support/attack (used on edges/relations from intermediate to output arguments, and on words from input to intermediate arguments).

PERCEPTRONS MULTICOUCHES ET RÉSEAUX CONVOLUTIFS

19

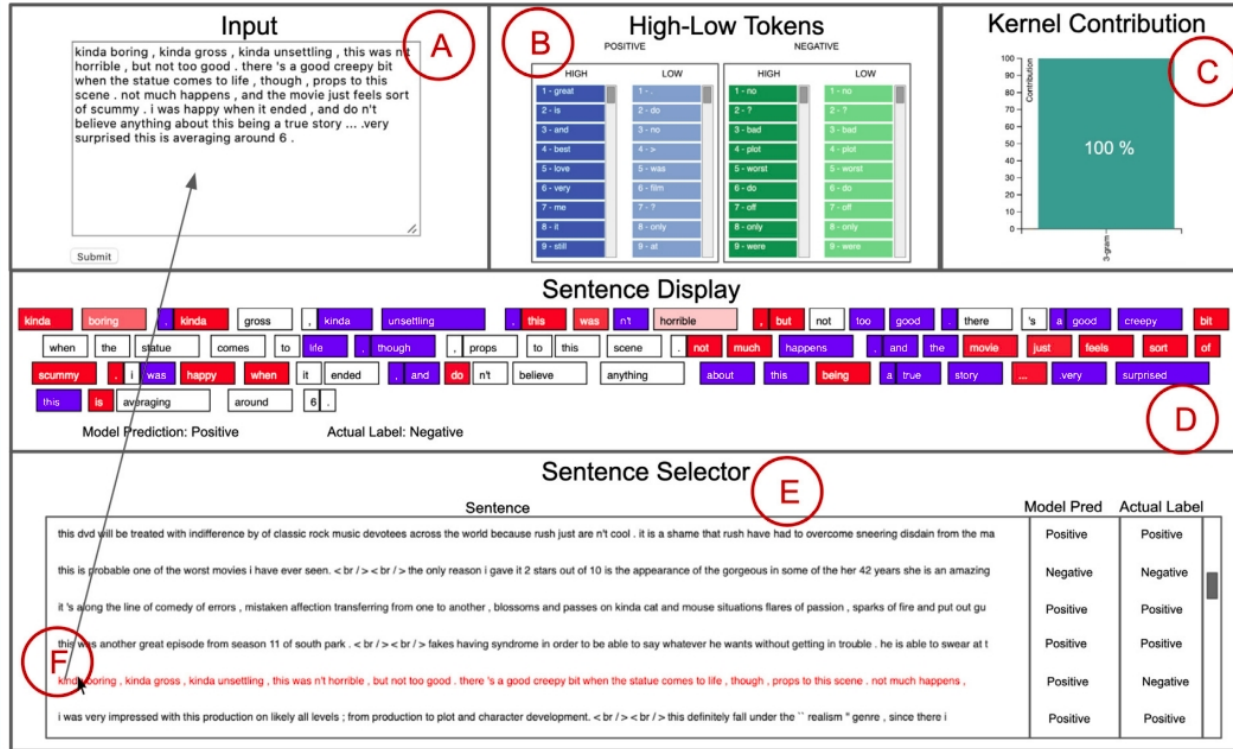


Fig. 6. Visual Analysis System.

Piyush Chawla, Subshashis Hazarika, and Han-Wei Shen. Token-wise sentiment decomposition for convnet : Visualizing a sentiment classifier. Visual Informatics, 4(2) :132-141, 2020.



PERCEPTRONS MULTICOUCHES ET RÉSEAUX CONVOLUTIFS

- Peu de travaux
- Les autres types de réseaux focalisent plus les recherches en explicabilité

RÉSEAUX DE NEURONES RÉCURRENTS

21



Fig. 1. The LSTMVis user interface. The user interactively *selects* a range of text specifying a hypothesis about the model in the Select View (a). This range is then used to *match* similar hidden state patterns displayed in the Match View (b). The selection is made by specifying a start-stop range in the text (c) and an activation threshold (t) which leads to a selection of hidden states (blue lines). The start-stop range can be further constrained using the pattern plot (d). The meta-tracks below depict extra information per word position like POS (e1) or the top K predictions (e2). The tool can then *match* this selection with similar hidden state patterns in the data set of varying lengths (f), providing insight into the representations learned by the model. The match view additionally includes user-defined meta-data encoded as heatmaps (g1,g2). The color of one heatmap (g2) can be mapped (h) to the word matrix (f) which allows the user to see patterns that lead to further refinement of the selection hypothesis. Navigation aids provide convenience (i_1, i_2).

Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M Rush. Lstmvis : A tool for visual analysis of hidden state dynamics in recurrent neural networks. IEEE TVCG, 24(1) :667–676, 2017.

RÉSEAUX DE NEURONES RÉCURRENTS

22

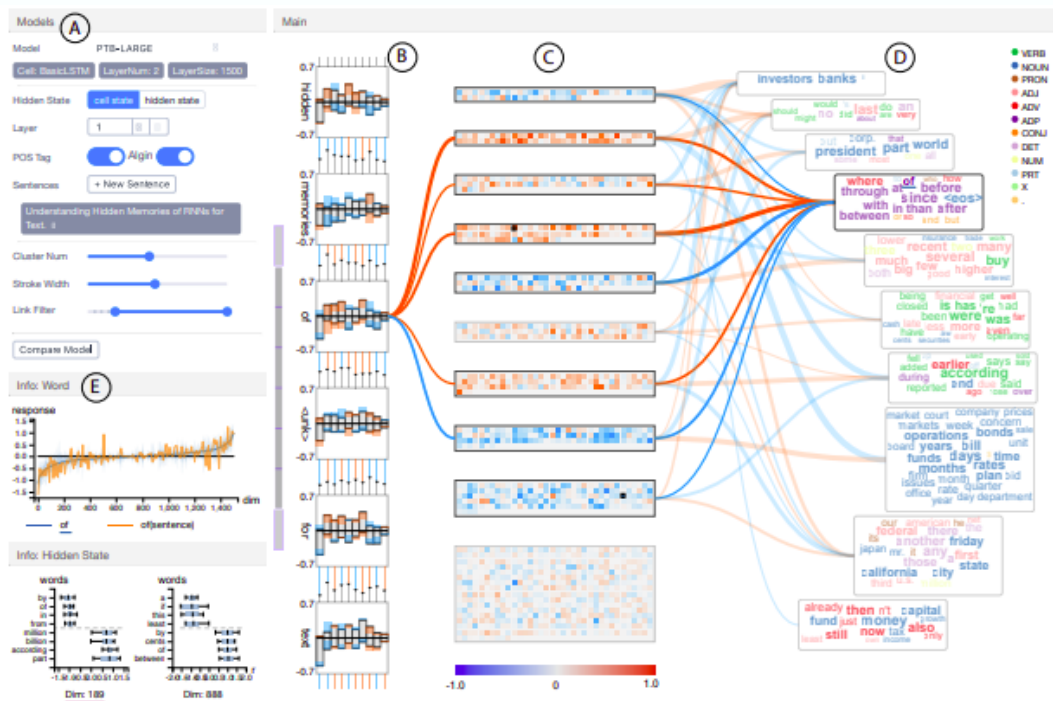


Figure 1: The interface of RNNVis. The control panel (A) shows parameter settings of an RNN and allows users to adjust visualization style. The main view (B-D) contains glyph-based sentence visualization (B), memory chips visualization for hidden state clusters (C), and word clouds visualization for word clusters (D). The detail view (E) shows the distributions of models responses to selected word "of" and interpretations of selected hidden units.

Yao Ming, Shaozu Cao, Ruixiang Zhang, Zhen Li, Yuanzhe Chen, Yangqiu Song, and Huamin Qu. Understanding hidden memories of recurrent neural networks. In 2017 IEEE VAST, pages 13–24. IEEE, 2017.

RÉSEAUX DE NEURONES RÉCURRENTS

23

Cell sensitive to position in line:

```
The sole importance of the crossing of the Berezina lies in the fact that it plainly and indubitably proved the fallacy of all the plans for cutting off the enemy's retreat and the soundness of the only possible line of action--the one Kutuzov and the general mass of the army demanded--namely, simply to follow the enemy up. The French crowd fled at a continually increasing speed and all its energy was directed to reaching its goal. It fled like a wounded animal and it was impossible to block its path. This was shown not so much by the arrangements it made for crossing as by what took place at the bridges. When the bridges broke down, unarmed soldiers, people from Moscow and women with children who were with the French transport, all--carried on by vis inertiae--pressed forward into boats and into the ice-covered water and did not surrender.
```

Cell that turns on inside quotes:

```
"You mean to imply that I have nothing to eat out of... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.
```

```
Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."
```

Cell that robustly activates inside if statements:

```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask, siginfo_t *info)
{
    int sig = next_signal(pending, mask);
    if (sig) {
        if (current->notifier) {
            if (sigismember(current->notifier_mask, sig)) {
                if (!current->notifier((current->notifier_data))) {
                    clear_thread_flag(TIF_SIGPENDING);
                    return 0;
                }
            }
        }
        collect_signal(sig, pending, info);
    }
    return sig;
}
```

A large portion of cells are not easily interpretable. Here is a typical example:

```
/* Unpack a filter field's string representation from user-space
 * buffer. */
char *audit_unpack_string(void **bufp, size_t *remain, size_t len)
{
    char *str;
    if (!*bufp || (len == 0) || (len > *remain))
        return ERR_PTR(-EINVAL);
    /* Of the currently implemented string fields, PATH_MAX
     * defines the longest valid length.
     */
```

Cell that turns on inside comments and quotes:

```
/* Duplicate LSM field information. The lsm_rule is opaque, so
 * it's uninitialized. */
static inline int audit_dupe_lsm_field(struct audit_field *df,
                                     struct audit_field *sf)
{
    int ret = 0;
    char *lsm_str;
    /* Our own copy of lsm_str. */
    lsm_str = kstrdup(sf->lsm_str, GFP_KERNEL);
    if (unlikely(!lsm_str))
        return -ENOMEM;
    df->lsm_str = lsm_str;
    /* Our own (refreshed) copy of lsm_rule. */
    ret = security_audit_rule_init(df->op, df->lsm_str,
                                   (void *)df->lsm_rule);
    /* Keep currently invalid fields around in case they
     * become valid after a policy reload. */
    if (ret == -EINVAL) {
        pr_warn("audit rule for LSM '%s' is invalid\n",
                df->lsm_str);
        ret = 0;
    }
    return ret;
}
```

Cell that is sensitive to the depth of an expression:

```
#ifdef CONFIG_AUDITSYSCALL
static inline int audit_match_class_bits(int class, u32 *mask)
{
    int i;
    if (classes[class]) {
        for (i = 0; i < AUDIT_BITMASK_SIZE; i++)
            if (mask[i] & classes[class][i])
                return 0;
    }
    return 1;
}
```

Cell that might be helpful in predicting a new line. Note that it only turns on for some "):

```
char *audit_unpack_string(void **bufp, size_t *remain, si
{
    char *str;
    if (!*bufp || (len == 0) || (len > *remain))
        return ERR_PTR(-EINVAL);
    /* Of the currently implemented string fields, PATH_MAX
     * defines the longest valid length.
     */
    if (len > PATH_MAX)
        return ERR_PTR(-ENAMETOOLONG);
    str = kcalloc(len + 1, GFP_KERNEL);
    if (unlikely(!str))
        return ERR_PTR(-ENOMEM);
    memcpy(str, *bufp, len);
    str[len] = 0;
    *bufp += len;
    *remain -= len;
    return str;
}
```

Figure 2: Several examples of cells with interpretable activations discovered in our best Linux Kernel and War and Peace LSTMs. Text color corresponds to $\tanh(c)$, where -1 is red and +1 is blue.

RÉSEAUX DE NEURONES RÉCURRENTS

24

true	predicted	N°	Notation: -- very negative, - negative, 0 neutral, + positive, ++ very positive
		1.	do n't waste your money .
		2.	neither funny nor suspenseful nor particularly well-drawn .
		3.	it 's not horrible , just horribly mediocre .
		4.	... too slow , too boring , and occasionally annoying .
		5.	it 's neither as romantic nor as thrilling as it should be .
	--	6.	the master of disaster - it 's a piece of druck disguised as comedy .
		7.	so stupid , so ill-conceived , so badly drawn , it created whole new levels of ugly .
		8.	a film so tedious that it is impossible to care whether that boast is true or not .
		9.	choppy editing and too many repetitive scenes spoil what could have been an important documentary about stand-up comedy .
	--	10.	this idea has lost its originality ... and neither star appears very excited at rehashing what was basically a one-joke picture .
		11.	ecks this one off your must-see list .
	-	12.	this is not a "friday" worth waiting for .
	-	13.	there is not an ounce of honesty in the entire production .
	-	14.	do n't expect any surprises in this checklist of teamwork cliché ...
	-	15.	he has not learnt that storytelling is what the movies are about .
	-	16.	but here 's the real damn : it is not funny , either .
	+	17.	these are names to remember , in order to avoid them in the future .
	-	18.	the cartoon that is n't really good enough to be on afternoon tv is now a movie that is not really good enough to be in theaters .
		19.	a worthy entry into a very difficult genre .
	++	20.	it 's a good film -- not a classic , but odd , entertaining and authentic .
	--	21.	it never fails to engage us .

Figure 2: LRP heatmaps of exemplary test sentences, using as target class the *true* sentence class. Positive relevance is mapped to red, negative to blue, and the color intensity is normalized to the maximum absolute relevance per sentence. The true sentence class, and the classifier's predicted class, are indicated on the left.

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining recurrent neural network predictions in sentiment analysis. In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 159–168, Copenhagen, Denmark, September 2017. ACL.



RÉSEAUX DE NEURONES RÉCURRENTS

25

- Activations, états cachés
- Carte de chaleur sur les mots
- Adaptation de mécanismes issu du traitement des images

RÉSEAUX DE NEURONES AUTO-ATTENTIFS

26

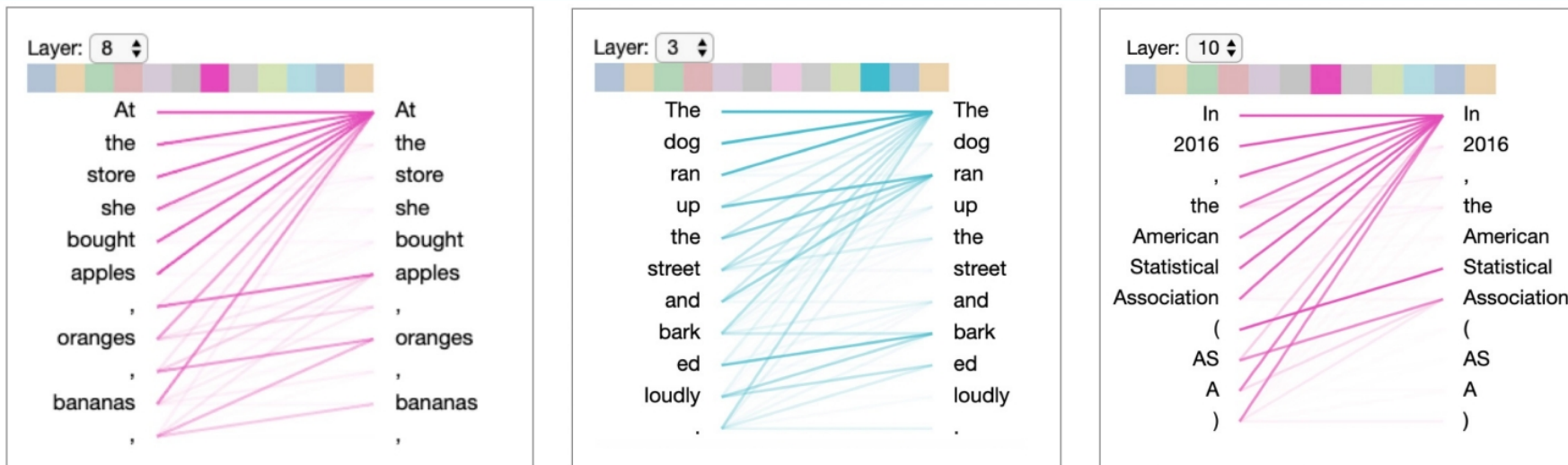


Figure 3: Examples of attention heads in GPT-2 that capture specific lexical patterns: list items (left); verbs (center); and acronyms (right). Similar patterns were observed in these attention heads for other inputs. Attention directed toward first token is likely null attention (Vig and Belinkov, 2019).

Jesse Vig. A multiscale visualization of attention in the transformer model. In 57th Annual Meeting of the Association for Computational Linguistics : System Demonstrations (ACL), pages 37-42. ACL, 2019.



RÉSEAUX DE NEURONES AUTO-ATTENTIFS

Débat autour de l'attention :

- Jain et al. : l'attention n'est pas une méthode d'explication
- Wiegrefe et al. : les travaux que les expérimentations de Jain et al. ne permettent pas de soutenir leur théorie.
- Ethayarajh et al. : il existe une grande similarité entre les représentations des mots dans les transformeurs (similarité cosinus). Ceci produit des matrices d'attention peu informatives.

RÉSEAUX DE NEURONES AUTO-ATTENTIFS

28

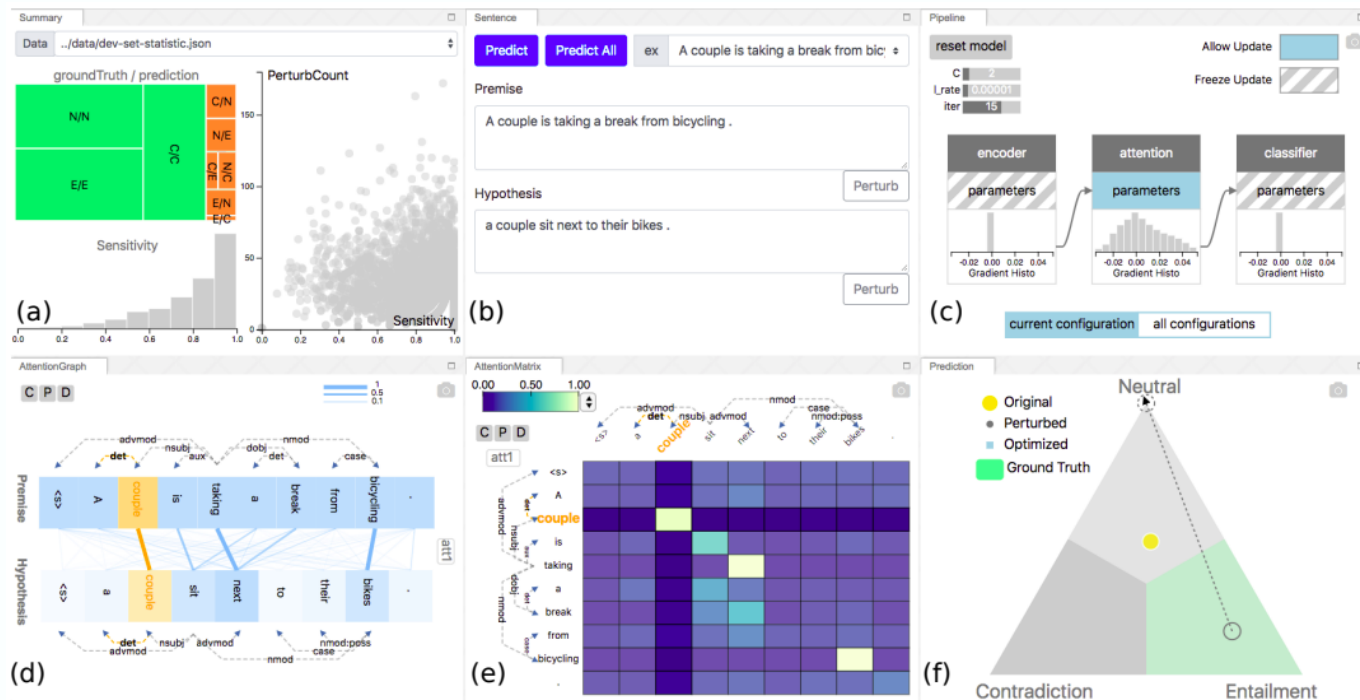


Fig. 1. The interface of the proposed system. During exploration, we can filter through a large number of sentence pairs summarized in (a). The current selected pair is displayed in (b). The model internal information (attention) is displayed in (d) and (e). The predicted probability (one of three labels: *neutral*, *entailment*, *contradiction*) is shown in the barycentric coordinate in (f). Finally, the high-level model structure and the updates to the model are summarized in (c).

Shusen Liu, Zhimin Li, Tao Li, Vivek Srikumar, Valerio Pascucci, and Peer-Timo Bremer. Nlize : A perturbation-driven visual interrogation tool for analyzing and interpreting natural language inference models. IEEE TVCG, 25(1) :651-660, 2018.



RÉSEAUX DE NEURONES AUTO-ATTENTIFS

- L'attention est centrale
- Nombreuses tâches de traitement automatique de langue

MÉTHODES AGNOSTIQUES

30

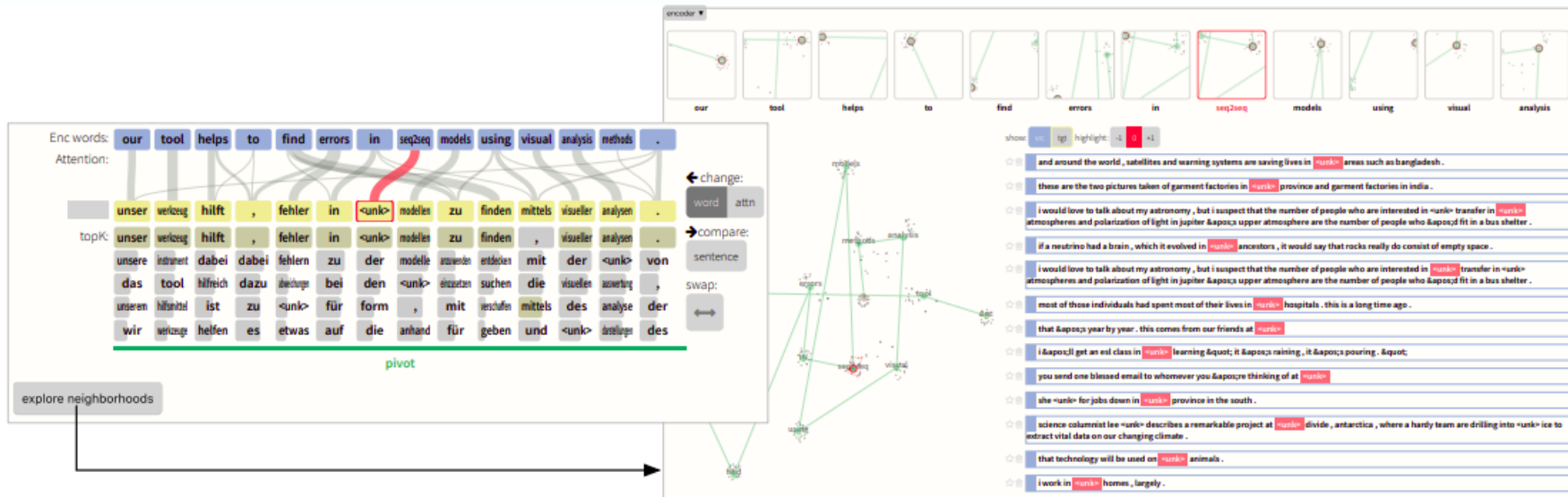


Fig. 1. Example of Seq2Seq-Vis. In the translation view (left), the source sequence “our tool helps to find errors in seq2seq models using visual analysis methods.” is translated into a German sentence. The word “seq2seq” has correct attention between encoder and decoder (red highlight) but is not part of the language dictionary. When investigating the encoder neighborhoods (right), the user sees that “seq2seq” is close to other unknown words “<unk>”. The buttons enable user interactions for deeper analysis.

Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, and Alexander M Rush. Seq2seq-vis : A visual debugging tool for sequence-to-sequence models. IEEE TVCG, 25(1) :353-363, 2018.

MÉTHODES AGNOSTIQUES

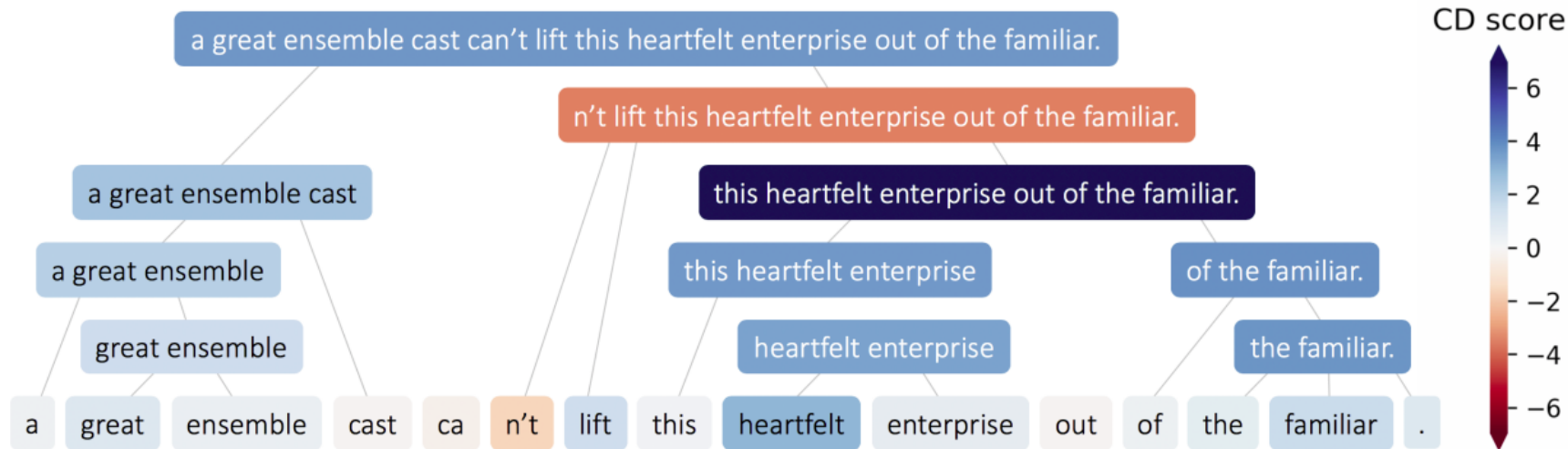


Figure 2: ACD interpretation of an LSTM predicting sentiment. Blue is positive sentiment, white is neutral, red is negative. The bottom row displays CD scores for individual words in the sentence. Higher rows display important phrases identified by ACD, along with their CD scores, converging to the model's (incorrect) prediction in the top row. (Best viewed in color)

MÉTHODES AGNOSTIQUES

32

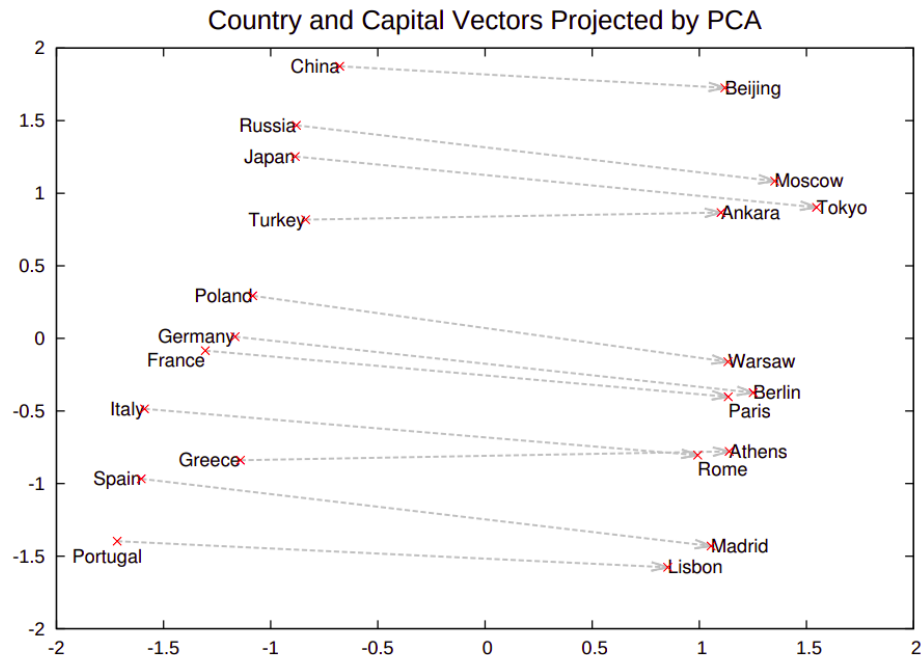


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

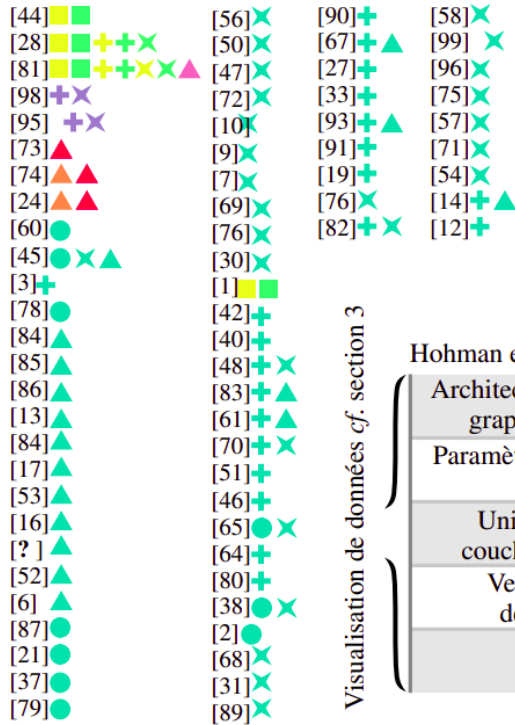


MÉTHODES AGNOSTIQUES

- Espace de représentation
- Explication hiérarchique plus récemment

VUE D'ENSEMBLE

34



Visualisation de données cf. section 3

		Interprétabilité cf. section 2							
		Transparence			Explication post-hoc				
		Compréhension globale du modèle	Compréhension des parties du modèle	Convergence vers une solution optimale	Explications verbales ou écrites	Explications locales	Explications de complexité modérée	Techniques de visualisation	
		Lipton [49]							
Hohman et al. [32]									
Architecture à l'aide de graphes des réseaux	■	4 articles	4 articles						Obj.1
Paramètres des réseaux de neurones	●							10 articles	Obj.2
Unités de calcul ou couches de neurones	+	2 articles	2 articles		2 articles			21 articles	Obj.3
Vecteurs et espace de représentation	✕	1 article	2 articles		2 articles			27 articles	Obj.4
Informations agrégées	▲			1 article		2 articles	3 articles	16 articles	Obj.5
		Transp.1	Transp.2	Transp.3	Exp.1	Exp.2	Exp.3	Exp.4	

TABLE 1 – Méthodes d’interprétabilité présentées selon les classifications de Hohman et al. [32] et de Lipton [49]. Les catégories ne sont pas exhaustives car, selon le sujet, certaines catégories se confondent (*i.e.*, les unités de calcul et les vecteurs de représentation dans le cas d’utilisation de *RNN* ou réseaux auto-attentionnels.)

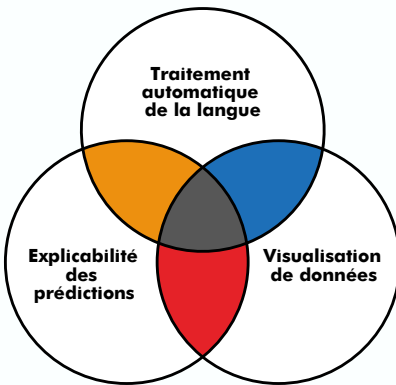


CONCLUSION

1. Visualisation gagne en intérêt, incontournable
2. Disparité des recherches en fonction des modèles
3. Il n'existe pas de terminologie fixée
4. Méthodes les plus utilisées :
 - Carte de chaleur
 - Graphe bipartie
 - Matrice d'attention
 - Espace de représentation

PERSPECTIVES DE NOS TRAVAUX

- Panorama → État de l'art



- Notions sciences sociales:
 - Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267, 1-38.



LIRMM



MERCI !



SIRIC
MONTPELLIER CANCER
Site de Recherche Intégrée sur le Cancer



MILLER, T. (2019)

- 1. Explanations are contrastive — they are sought in response to particular counterfactual cases, which are termed foils in this paper. That is, people do not ask why event P happened, but rather why event P happened instead of some event Q. This has important social and computational consequences for explainable AI.
- 2. Explanations are selected (in a biased manner) — people rarely, if ever, expect an explanation that consists of an actual and complete cause of an event. Humans are adept at selecting one or two causes from a sometimes infinite number of causes to be the explanation. However, this selection is influenced by certain cognitive biases.
- 3. Probabilities probably don't matter — while truth and likelihood are important in explanation and probabilities really do matter, referring to probabilities or statistical relationships in explanation is not as effective as referring to causes. The most likely explanation is not always the best explanation for a person, and importantly, using statistical generalisations to explain why events occur is unsatisfying, unless accompanied by an underlying causal explanation for the generalisation itself.
- 4. Explanations are social — they are a transfer of knowledge, presented as part of a conversation or interaction, and are thus presented relative to the explainer's beliefs about the explainee's beliefs.



ANDREAS MADSEN (2022)

- Terminology
- Synergy
- Helpful complex models