TRAITEMENT AUTOMATIQUE DE LA LANGUE: PLONGEON AU CŒUR DES WORD EMBEDDINGS

De Word2vec à BERT, un espoir nouveau pour le TAL

Mikaël Dusenne

2021-11-23

INTRODUCTION

80% des données cliniques pertinentes sont non structurées

Comment représenter efficacement les données textuelles pour l'apprentissage automatique?

Apprentissage automatique et langage naturel

Approches classiques : un mot / n-gram = une variable

Problèmes :

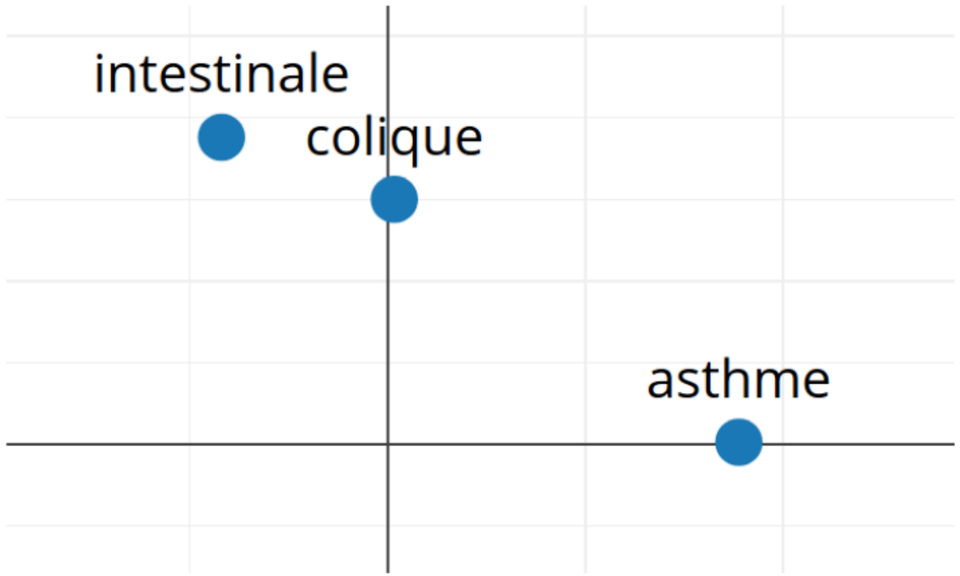
- Pas de notion de **distance sémantique**
- Très grand **nombre de variables**
- Données **éparses**

mot	hospitalisé	asthme	occlusion	...	colique	intestinale	aigüe
asthme	0	**1**	0	...	0	0	0
colique	0	0	0	...	**1**	0	0
intestinale	0	0	0	...	0	**1**	0

Word Embeddings

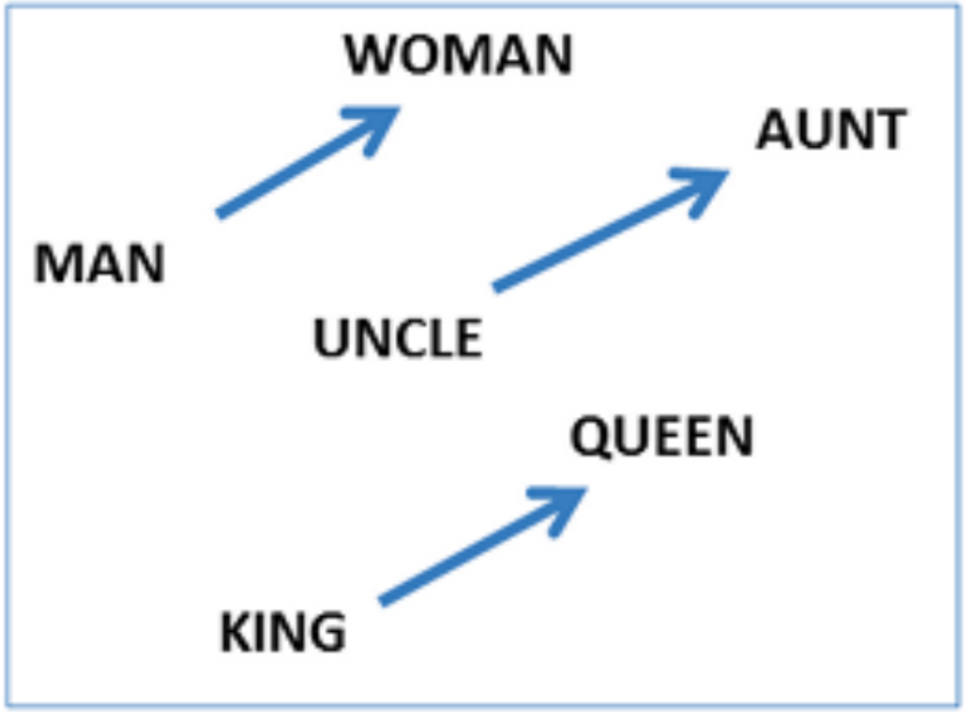
- Représentation **dense** des mots
- Vecteurs de **nombres réels**
- Dimension **indépendante de la taille du vocabulaire**
- Proximité dans l'espace vectoriel corrélée à la **similarité sémantique**

mots	0	1
asthme	0.888	0.014
colique	0.017	1.500
intestinale	-0.420	1.880



Word Embeddings

Les Embeddings permettent d'utiliser le calcul vectoriel pour effectuer des transformations sémantiques



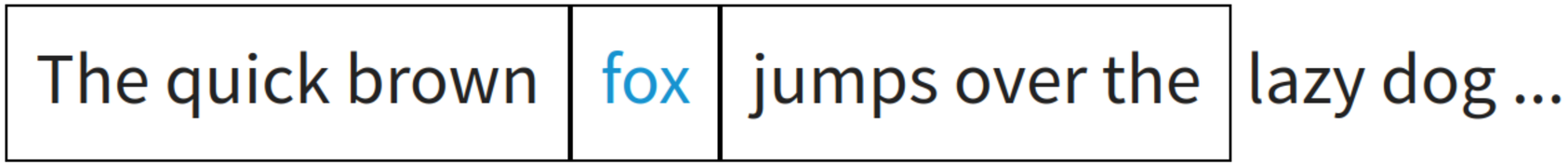
King + (Woman - Man) = Queen

Embeddings et TAL : implémentations

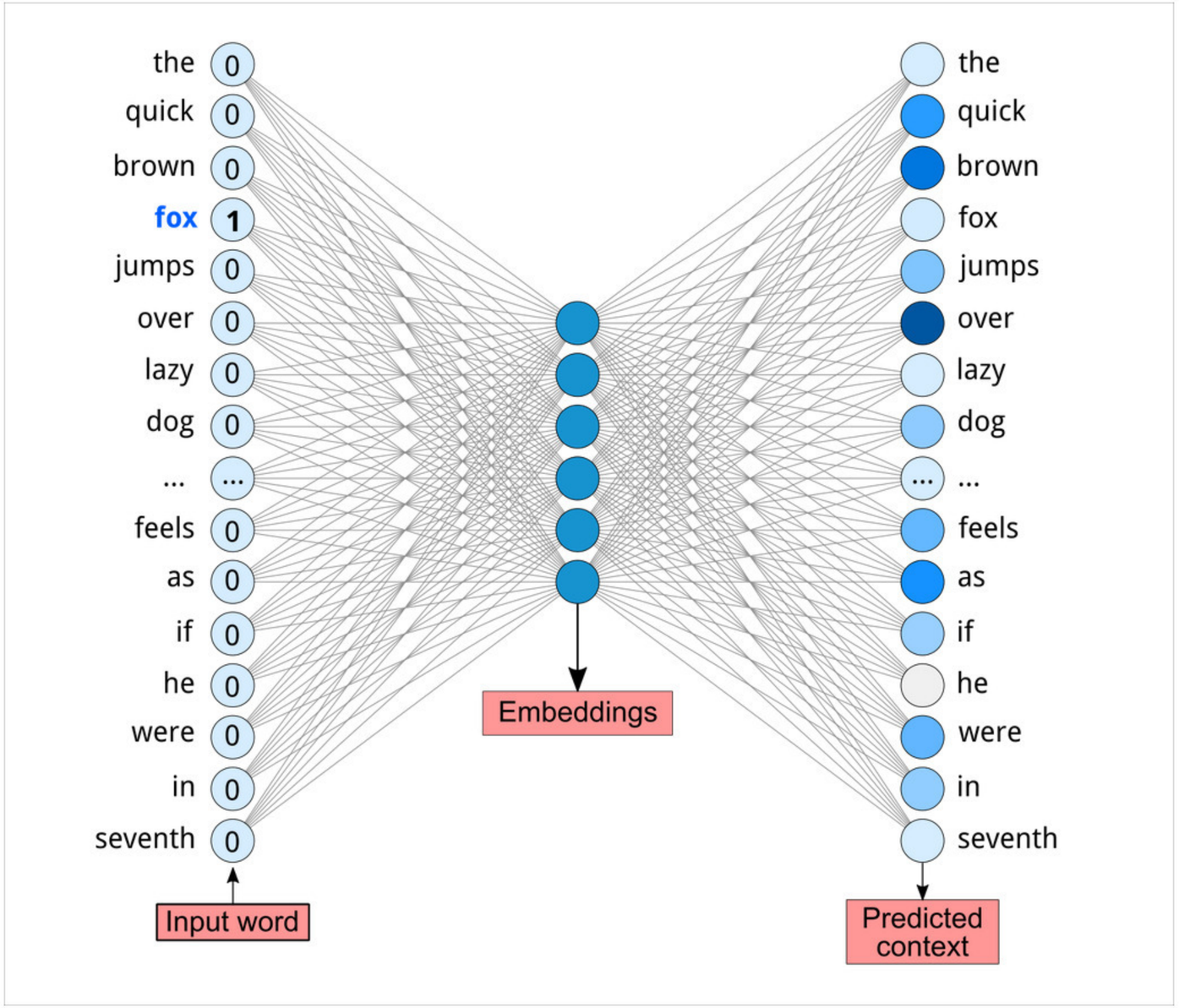
- **2013** : Word2Vec 2013[1] :
  - réseau de neurones pour créer les embeddings
- **2014** : GloVe [2]
  - "global vectors", matrice de co-occurence utilisant le corpus entier
- **2014** : Doc2Vec [3]
  - Vecteurs de Documents
- **2016** : FastText [4]
  - Décomposition des mots en n-grams de caractères
- **2018** : ELMo [5]
  - utilise l'ordre des mots (LSTM bi-directionnel)
- **2018** : BERT [6]
  - utilise des "attention network" (Transformer)
  - gestion des homonymes
- **2018** : Flair [7]
  - Zalando Research
  - Étiquetage morpho-syntaxique
- **2019** : ALBERT [8]
  - Améliore BERT : moins de paramètres, entraînement plus rapide
- **2019** : BioBERT [9]
  - pré-entraîné sur pubmed et PMC (en anglais)
- **2019** : camemBERT [10]
  - pré-entraîné sur un corpus français (OSCAR corpus, non médical)
- **2019** : FlauBERT [11]
  - pré-entraîné sur un corpus français (non médical)

2013 - word2vec

- Première adaptation réellement fonctionnelle des techniques d'embedding au TAL
- réseau de neurones simple
- apprentissage "semi-supervisé"



2013 - word2vec



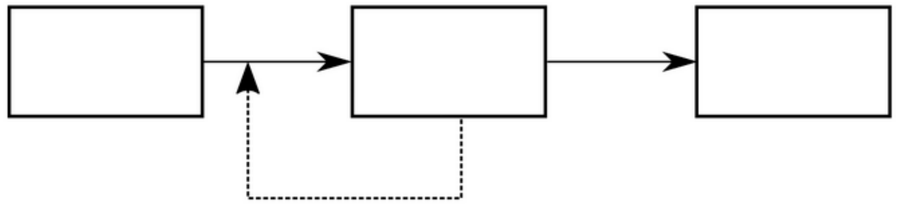
2013 - word2vec

Inconvénients:

- Ne prend pas en compte l'ordre des mots dans le contexte
- un mot a toujours la même représentation -> problème pour la polysémie

2015 - Représentations contextuelles

Besoin d'une solution pour prendre en compte l'ordre des mots



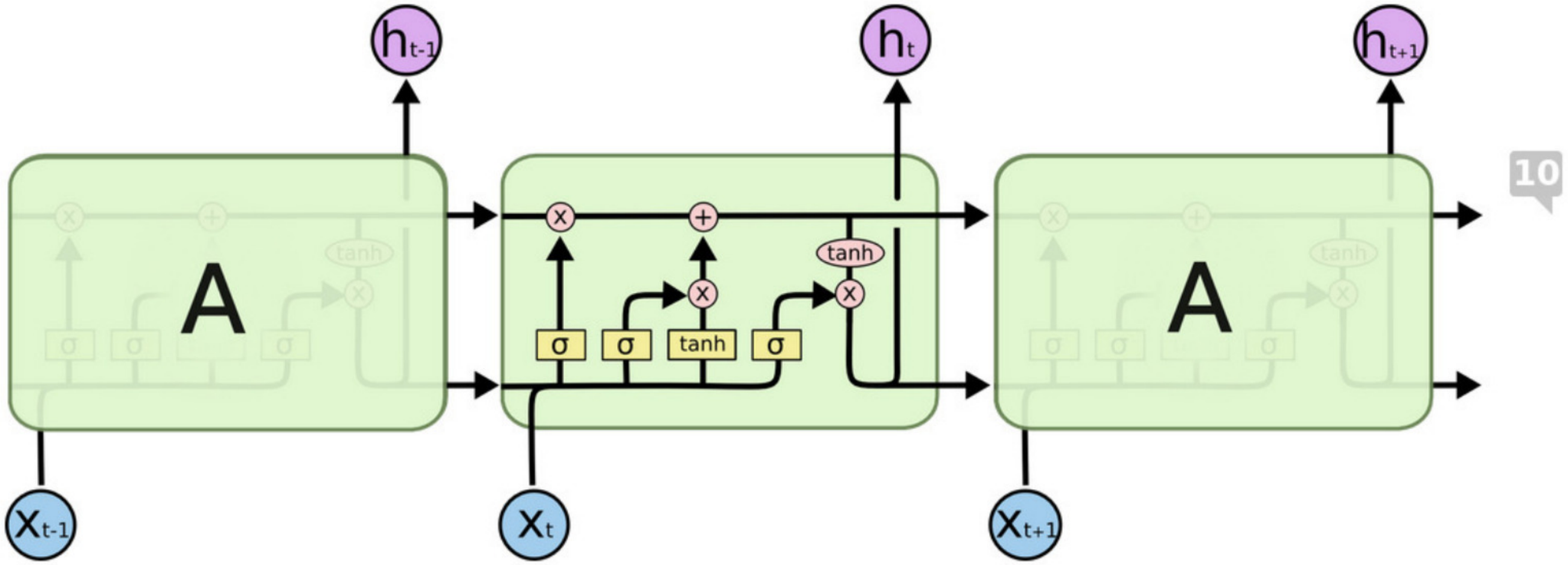
Le LSTM est un type de réseau neuronal récurrent qui introduit des boucles permettant à une information apparue précédemment d'être "mémorisée".

-> Respect de l'ordre d'apparition des mots dans une phrase.

Semi-supervised Sequence Learning (Andrew M. Dai, Quoc V. Le)

2015 - Représentations contextuelles

Besoin d'une solution pour prendre en compte l'ordre des mots

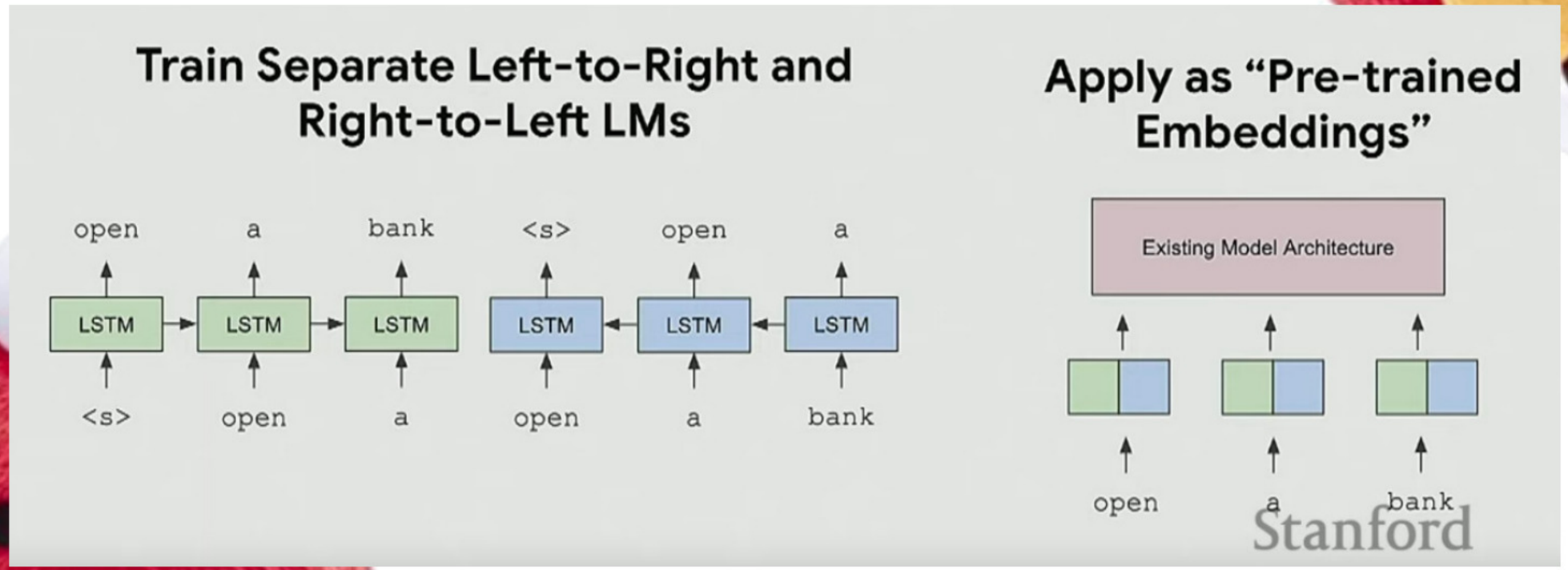


The repeating module in an LSTM contains four interacting layers.

2017 - ELMo

EMBEDDINGS FOR LANGUAGE MODELS (ELMO)

Combine un LSTM "en avant" et un LSTM "en arrière"

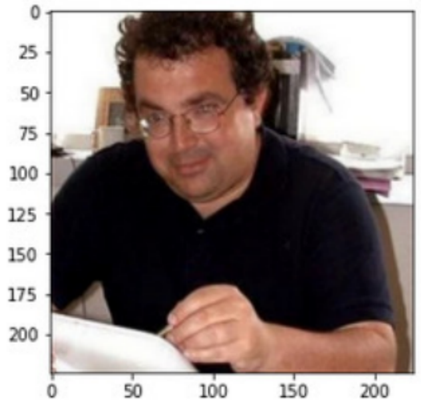


source: Stanford CS224N: NLP with Deep Learning | Winter 2020 | BERT and Other Pre-trained Language Models

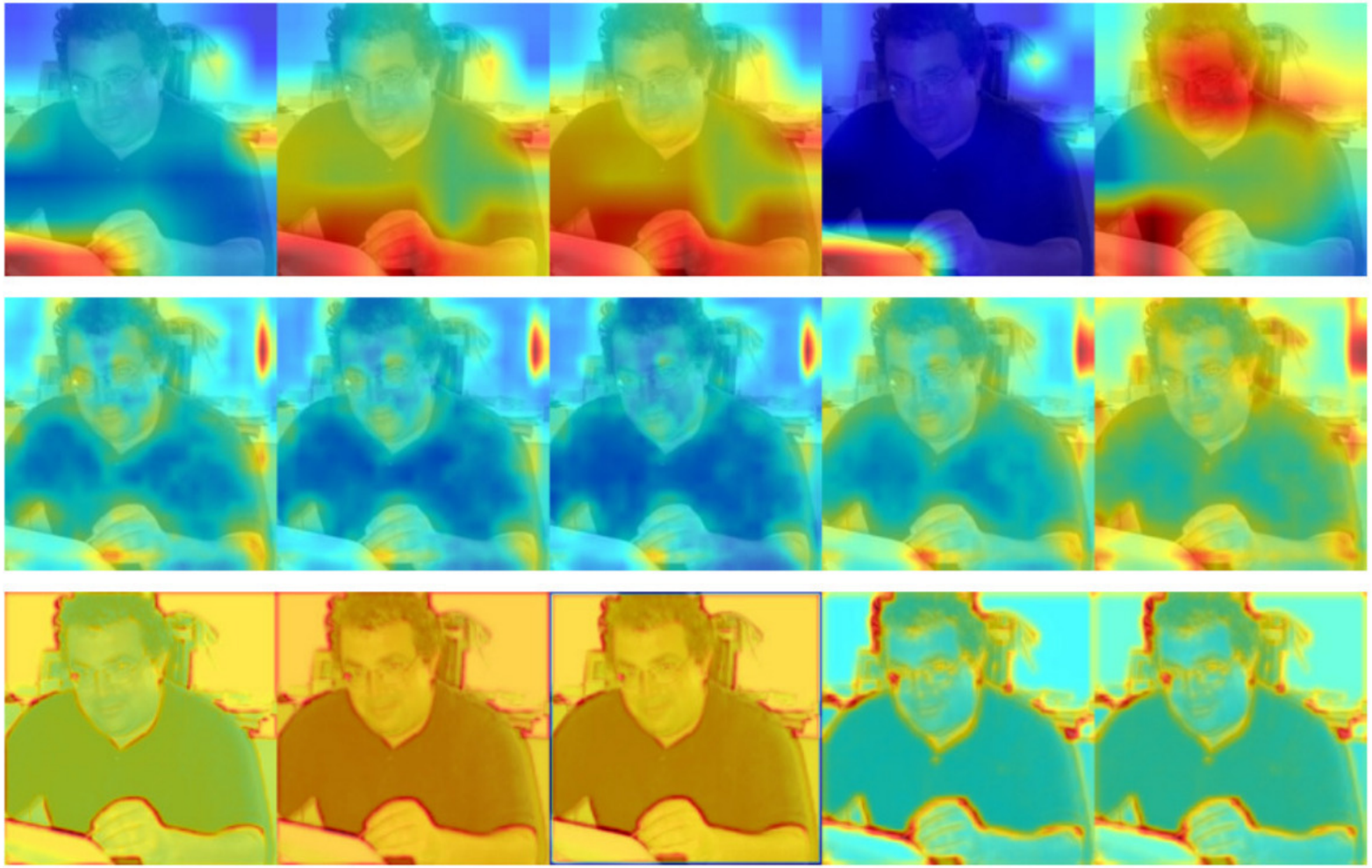
2018 - Generative Pre-trained Transformer

- GPT1 créé par OpenAI
- Utilise pour la première fois les mécanismes d'Attention apportés par les transformers pour accorder une importance plus grande à certains mots
- Publication: Attention Is All You Need (Ashish Vaswani et al.)

Réseaux neuronaux et Mécanisme d'Attention



Réseaux neuronaux et Mécanisme d'Attention



Réseaux neuronaux et Mécanisme d'Attention

Transformers

- Les transformers permettent d'utiliser les mécanismes d'attention afin de se référer aux mots précédents d'une phrase, sans avoir recours aux réseaux récurrents (LSTM).
- utilisent le matériel récent (GPU, TPU) de façon beaucoup plus efficace (parallélisation plus simple)

Inconvénients de GPT

Le modèle est uni-directionnel (le texte n'est lu que de gauche à droite)

2018 - Bidirectional Encoder Representations from Transformers (BERT)

- Réellement bidirectionnel
- Utilise les transformers
- entraînement en [MASK]ant 15% des mots de chaque phrase

Performances

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

Table 1: GLUE Test results, scored by the evaluation server (https://gluebenchmark.com/leaderboard). The number below each task denotes the number of training examples. The "Average" column is slightly different than the official GLUE score, since we exclude the problematic WNLI set.[8] BERT and OpenAI GPT are single-model, single task. F1 scores are reported for QQP and MRPC, Spearman correlations are reported for STS-B, and accuracy scores are reported for the other tasks. We exclude entries that use BERT as one of their components.

BERT-like

- (2019) BioBERT : a pre-trained biomedical language representation model for biomedical text mining
  - BERT entraîné sur des documents biomédicaux en anglais
  - amélioration significative des performances sur des tâches biomédicales
  - démontre la possibilité de créer des modèles spécialisés dans certains domaines
- (2019) RoBERTa : A Robustly Optimized BERT Pretraining Approach
  - Réévalue et améliore l'entraînement de BERT
- (2019) ALBERT : A Lite BERT for Self-supervised Learning of Language Representations
  - optimisation de BERT: réduction drastique du # de paramètres (12M, -89%)
- (2019) StructBERT : Incorporating Language Structures into Pre-training for Deep Language Understanding
  - focus sur la structure du langage, ajout d'une tâche de reconstruction de l'ordre des mots / phrases pendant l'entraînement
- (2019) TinyBERT : Distilling BERT for Natural Language Understanding
  - 7.5x smaller, 9.4x faster, 96.8% of BERT performances on GLUE
- (2019) FlauBERT : Unsupervised Language Model Pre-training for French
  - entraîné sur l'ordinateur Jean Zey au CNRS (28 PétaFlops)
  - sur un corpus français généraliste
  - FLUE
- (2020) DeBERTa : Decoding-enhanced BERT with Disentangled Attention
  - amélioration de la gestion de la position des mots
- (2021) BERTAC : Enhancing Transformer-based Language Models with Adversarially Pretrained Convolutional Neural Networks
  - CNN utilisant un apprentissage de type GAN sur le texte de wikipedia, puis combiné à ALBERT
- (2020) CamemBERT : a Tasty French Language Model
  - Basé sur RoBERTa
  - entrainé sur le corpus multilingue OSCAR

TRAVAUX AU D2IM

Travaux Emeric Dynomant

- Sujet : Bioinformatics articles structuring with an end-to-end processing pipeline
- Machine Learning for NLP; word & document embeddings for text
- Word embeddings
  - Comparaison de cinq algorithmes sur 11,8 M de documents de santé d'un EDS
- Document embeddings
  - Doc2Vec2PubMed vs. algorithme actuel Related Articles

2019 Jul 29;7(3):e12510.

Medical word embeddings querying page

12M | endocardite | Search

GloVe	infectieuse, myocardite, eto, native, streptocoque, bovis, bactériémie, faecalis, _endocardite
FastText (CBOW)	proprio_septive, myopericardite, septo_optique, endo_aortique, endocartite, endoculaire, acrodermite, rhino_septale, salmonellose, épidermolyse
Word2Vec (Skip-Gram)	bovis, sanguinis, eto, gordonii, gallolyticus, aorto_mitrale, mutans, infectieuse, streptoccoque, salivarius
FastText (Skip-Gram)	endocartite, endocardique, proctologique, extancilline, septo_basale, prolongements, recanalisée, précentrale, dantrolene, podoscopique
Word2Vec (CBOW)	endocartite, _endocardite, native, bovis, médiastinite, myocardite, mutans, gallolyticus, myopéricardite, tamponnade

Home

Word embeddings dans deux contextes différents

QUERY : "facebook"

LiSSa corpus (300k)	internet, twitter, web, blog, e_learning, blogs, internautes, tic, game, …
RUH documents (12M)	reproches, injures, messages, insultes, rumeurs, ex_conjointe, menaces, insultant, …

Espace vectoriel disponible pour la communauté scientifique

Annotateur Sémantique

Intégration de BERT* pour améliorer l'annotateur sémantique de l'EDSaN ?

MERCI