On the Behaviour of BERT's Attention for the Classification of Medical Reports

Luca Putelli[1], Alfonso E. Gerevini[1], Alberto Lavelli[2], Tahir Mehmood[1] and Ivan Serina[1]

[1]*Università degli Studi di Brescia, Brescia, Italy*
[2]*Fondazione Bruno Kessler, Povo (TN), Italy*

Abstract

Since BERT and the other Transformer-based models have been proved successful in many NLP tasks, several studies have been conducted to understand why these complex deep learning architectures are able to reach such remarkable results. Such studies have focused on visualising and analysing the behaviour of each self-attention mechanism and are often conducted with large, generic and annotated datasets for the English language, using supervised probing tasks in order to test specific linguistic capabilities. However, in several practical contexts there are some difficulties: probing tasks may not be available, the documents can contain a strict technical lexicon, and the datasets can be noisy. In this work we analyse the behaviour of BERT in a specific context, i.e. the classification of radiology reports collected from an Italian hospital. We propose (i) a simplified way to classify head patterns without relying on probing tasks or manual observations, and (ii) an algorithm for extracting the most relevant relations among words captured by each self-attention. Combining these techniques with manual observations, we present several examples of linguistic information that can be extracted from BERT in our application.

1. Introduction

Language models based on Transformer [1] like BERT (Bidirectional Encoder Representations from Transformer) [2], have obtained remarkable results in Natural Language Processing (NLP) tasks like machine translation or text classification, reaching a new state of the art. Intuitively, these models are composed by several encoders that progressively learn information about words and how they are related to each other. Encoders are made (among other components) by several parallel self-attention mechanisms called *heads*. For each word, a head calculates a probability distribution representing how much this word is related to every other word contained in the document.

The results and the complexity of BERT lead several research groups [3] to study how these language models capture the structure of the language [4], grammatical knowledge [5, 6, 7] or task specific information [8]. These analyses are conducted focusing on the embedded representation provided by each encoder [9] or on the self-attention mechanism in each head [10, 11, 12]. Similarly to other deep learning techniques such as LSTM Neural Networks [13],

XAI.it 2022 - Italian Workshop on Explainable Artificial Intelligence
luca.putelli1@unibs.it (L. Putelli); alfonso.gerevini@unibs.it (A. E. Gerevini); lavelli@fbk.eu (A. Lavelli); tahir.mehmood@unibs.it (T. Mehmood); ivan.serina@unibs.it (I. Serina)

BERT has been proven effective for extracting information from clinical narrative texts [14, 15]. Therefore, this technology could improve the efficacy and quality of patients' care. However, in this environment it is very important to assure the physicians that the system is correct and to expose the reasoning behind its decisions [16].

Moreover, according to works like [13, 17, 18], the Attention Mechanism can be used for highlighting the most important sections in a document and exploit them for interpretability, simply extracting the weights assigned by the Attention Mechanism, which can be seen as an indicator of the importance of each word for the predictive task. However, the self-attention used in BERT assigns a weight representing how much two words are related to each other. Moreover, while in LSTM-based models there is usually a single attention mechanism, in BERT there are more than 100 heads to consider. Therefore, deriving a single and straightforward indication of the importance of words for the classification task is definitely more challenging.

Furthermore, the studies in [10, 12] show that heads can be grouped according to a few distinct patterns; for instance, there are heads that always connect a word with the previous one or that distribute the attention weights across several words. This grouping process can be done by inspecting the heads manually [11] or with clustering techniques [19]. Both alternatives present some issues, such as the necessity of human intervention if the inspection is done manually, while in the second alternative the variability of the results depends on which clustering algorithm is selected and its hyper-parameters. These works usually show and verify the head behaviour on benchmark datasets in English and they use probing tasks, i.e., supervised classification tasks that focus on the capability of the self-attention weights to encode linguistic knowledge, without explicitly extracting the connection between words with the highest weights. This is not a trivial task given the differences among the heads.

In this work, we apply BERT in the context of the classification of radiology reports written in Italian and collected from the radiology department of *Spedali Civili di Brescia*. Then, we analyse the behaviour of BERT's Attention, presenting a schematic grouping process for the heads that does not require human inspection or clustering algorithms. Moreover, we propose an algorithm for extracting the most important word pairs according to the self-attention weights provided by each head. We then verify how these procedures can be exploited in our context for extracting useful information and their relation to the interpretability of BERT.

2. Background and Related work

2.1. BERT

BERT [2] is an architecture based on Transformer [1] composed by several encoding layers which progressively analyse a sequence of tokens (i.e., words or parts of a word) in order to capture their meaning. Each layer applies multiple self-attention mechanisms (called *heads*) in parallel. Considering a sequence of tokens S of length N, this mechanism produces a matrix $A_{i,j} \in \mathbb{R}^{N \times N}$, where i is the number of the encoding layer and j is the head number. For each token $w \in S$, the vector $a_w \in A_{i,j}$ contains the attention weights that represent how much w is related to the other tokens in S.

In order to calculate these weights, in each head the input representation of the token sequence $X \in \mathbb{R}^{N \times d}$ is projected into three new representations called key (K), query (Q) and value (V)

with three matrices W_k, W_q and W_v:

$$K = X \times W_k,\ Q = X \times W_q,\ V = X \times W_v \tag{1}$$

Then, the attention weights are calculated using a scaled dot-product between Q and K and applying the softmax function. The new token representation Z is calculated by multiplying the attention weights for V.

$$A = softmax(\frac{Q \times K}{\sqrt{d}}),\ Z = A \times V \tag{2}$$

where d is the length of the input representation of each token. Given that in each encoding layer there are multiple heads, in order to create a representation provided by the *multi-head attention mechanism* the result of each head is concatenated and then fed to a feed-forward layer. As described in [1], the multi-head attention mechanism is followed by a feed-forward layer and residual connections. The output of an encoding layer is the input of the next one.

Exploiting a large collection of documents, BERT is trained for two tasks: *language modeling*, where BERT learns to predict a percentage (usually 15%) of tokens from context, and *next sentence prediction*, which is a binary classification task where BERT has to predict if a sequence of two sentences is correct or not. For the latter task, BERT introduces two special tokens: [CLS], whose representation is used for the binary classification task and represents the whole sequence, and [SEP] which separates the two sentences. Learning these two tasks allows BERT to create a meaningful representation of each token and also to summarise the most important information in a sentence. Once the model is trained, it can be adapted using smaller datasets for specific NLP tasks like Named Entity Recognition, text classification, sentiment analysis, etc.

2.2. Related work

In the last few years, several studies have been conducted in order to understand the reasons behind its success and which linguistic and world knowledge is stored in a BERT model [3]. We can group these studies into two main categories, simply considering if they analyse the embedding representations or the head's behaviour. From the former category, the works in [6, 20] show how linguistic information (a total of 68 features such as Part-of-speech tags, verb inflection, depth of the dependency tree, etc.) is encoded in the BERT embeddings produced by each layer. Testing semantic roles, semantic dependencies (such as coreference between nouns and pronouns), entities and relations like in the classical NLP pipeline is the focus of [7]. Similar tests were conducted in [5] and [4], with a particular focus on subject-verb agreement. An aspect that regards all these works is that they require datasets annotated with the linguistic phenomena they are examining with their probing tasks. Unfortunately, in the clinical domain it is much more difficult to find this kind of annotations [21], and there are many challenges related to the quality of the text, with abbreviations, typos and ungrammatical language [16]. Therefore the use of probing tasks is quite limited, especially for the Italian language.

The latter category regards the analysis of the head's behaviour. Since the introduction of the first visualization tools, like BertViz [22], it has been possible to note how heads behave accordingly to some recognizable patterns. The work in [10] presents some interesting results,

manually selecting heads that give attention broadly, to the next token, or to [SEP]. The authors use probing tasks in order to show that certain heads target specific linguistic information, such as coreference, direct objects, relations between possessive pronouns and nouns, etc. In [11], similar patterns were presented and other probing tasks were executed. For instance, given a pair of tokens with a specific linguistic relation, they detect which heads assign a high weight to such pair. Our work differs from these ones because we propose a way of grouping heads according to their pattern with a quantitative approach. Moreover, while these works use probing tasks in order to find meaningful relations between pair of tokens without explicitly extracting them, we propose an algorithm specifically designed for this extraction, simplifying the subsequent analyses. In [19], the authors exploit clustering algorithms to automatically group heads into categories, and discuss their importance. A drawback of this approach is that clustering algorithms can produce very different results according to their implementation, their hyper-parameters and the number of selected clusters.

Evaluating the interpretability of deep learning systems for NLP, in terms of highlighting the most important snippets in order to justify the model output, is a very active research field [23]. While several studies focused on whether the attention weights could lead or not to an insight of the reasoning of the model [24, 25, 26], the research regarding BERT is still ongoing. In [27], the embedding representations are analysed in the context of a Question Answering task. In [8], the authors show that the words receiving most of the attention belong to the specific lexicon of the document subject. We have performed the same analysis in our context but with no results. In our opinion, this may depend on the fact that our radiology reports mostly share the same lexicon, with just small differences.

3. Methodology

In this section, we show the techniques used for identifying the heads' pattern and for extracting the most relevant relations between words, according to the attention weights distribution. Our goal is to find interesting information encoded in the attention weights for each head. However, given that this information is not labelled in our corpus of reports, first we want to assess the behaviour of each head, potentially selecting the most promising ones. Therefore, first we propose a simple method for identify the behaviour of each head and to group them accordingly. Next, we propose an algorithm for extracting the relation between word pairs with the highest weights that works regardless the difference between the heads' behaviours.

3.1. Metrics for the Head Grouping

Given a document made by N tokens, as described in Section 2.1, the head (i, j), where i is the number of the encoding layer and j is the head number in its multi-head self-attention mechanism, we call $A_{i,j} \in \mathbb{R}^{N \times N}$ the matrix of the attention weights produced by (i, j). For each token w, $A_{i,j}$ contains a vector $a_w \in \mathbb{R}^N$ which is a probability distribution representing its connections with all N tokens (itself included).

As reported in [10], when the attention is only on the special token [SEP], which is not used in the classification process but only for marking the end of the document, it can be seen as a null operation, or *no-op*. Therefore, first of all we want to evaluate how much a_w is close to a

no-op. Ideally, if all the attention is directed to [SEP], the probability distribution of the weights is a one-hot vector where the 1 is present in the index of [SEP]. In Equation 3, we refer to it as O. In order to calculate how much a_w focuses on tokens different from [SEP], we calculate the *No-Op Metric* ν as:

$$\nu_w = JSD(a_w||O),\ O[k] = \begin{cases} 0 & T[k] \neq [SEP] \\ 1 & T[k] = [SEP] \end{cases} \tag{3}$$

where JSD is the Jensen-Shannon Divergence, which is a standard statistical method for evaluating the similarity between two probability distributions. The Jensen-Shannon Divergence is bounded between 0 and 1, and it is 0 when the two distributions are identical, 1 when they are completely different. Please note that this metric is not calculated for evaluating the behaviour of the entire head, instead it is designed for a single token. We can describe a token w as *operative* if $\nu_w > 0.5$, otherwise we call it *not operative*. Using this metric, we can introduce a first categorization of the heads. As reported in [10], if a head executes a specific function, such as connecting a verb with its direct object, then those tokens onto which the function cannot be applied are usually connected to [SEP]. Therefore, we can specify two categories of head patterns: **General** if more than 50% tokens are operative and **Mixed** otherwise.

Considering the General heads, if the attention weights are distributed uniformly across the tokens, then no particular information could be extracted from them. Therefore, we evaluate how much a_w is similar with respect to a standard uniform distribution, using the *Focus Metric* ϵ, which we can define as:

$$\epsilon_w = JSD(a_w||U),\ U[k] = \frac{1}{N} \forall k \in [1, N] \tag{4}$$

Moreover, in several examples, we have observed a small number of heads where most tokens basically give high attention to themselves. Given their peculiar behaviour, we design a specific metric for identifying them: the *Self Metric* σ. We evaluate the difference between a_w and a one-hot vector, where the 1 is in the same position of w in the document. More formally,

$$\sigma_w = JSD(a_w||S_w),\ S_w[k] = \begin{cases} 0 & S_w[k] \neq w \\ 1 & S_w[k] = w \end{cases} \tag{5}$$

In order to capture the behaviour of each head, we calculate the average of ϵ and σ across all the tokens in a document. Observing the pattern varying the average of ϵ, we can group the general heads into four sub-categories:

- **Broadcast**, with an average ϵ lower than 0.4. These heads distribute their attention broadly across all tokens, with no particular criteria. This group resembles the *Dense* cluster described in [19];
- **Offset**, with an average ϵ higher than 0.7. These heads usually focus their attention on the previous/subsequent tokens without specific linguistic patterns. This group strongly resembles the *Diagonal* group observed in [11];
- **Local**, with an average ϵ between 0.4 and 0.7, and σ below 0.6. These heads mostly give attention to other tokens in the same sentence, with variable distance and behaviour depending on the analysed token. While some of these heads can be associated with

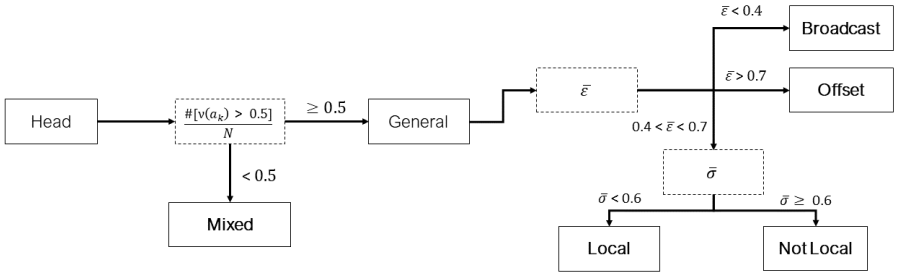


Figure 1: Diagram explaining the process of head categorisation. $\bar{\epsilon}$ stands for the average ϵ across all tokens, while $\bar{\sigma}$ stands for the average σ.

the *Block* group defined in [11] and the *Dense&Vertical* group defined in [19], a strong correlation with other groups has not been observed;

- **Not Local**, with an average ϵ between 0.4 and 0.7, and σ higher than 0.6. These heads give attention mostly to the token itself, other occurrences of the same word or other similar words, regardless if they are in the same sentence or not. To the best of our knowledge, this pattern is not present in the literature. A more detailed description, with examples and numerical results taken from our case study, can be seen in Section 5.2.

For the Mixed heads, given that the majority of the tokens is connected to [SEP], discovering which function they are trying to implement is a much more difficult process without probing tasks [10] and a general classification has not been proposed yet. For instance, the works in [19] and [11] group them into a category called Vertical, which basically highlights that the majority of tokens are connected to [SEP] with no further analysis. In Figure 1, we show the main flow of our categorisation process. Although our metrics are calculated for a single document, we have observed that there are no important differences in the behaviour of a head for different documents.

An important note about the thresholds used for the described metrics is that they have been set with a bottom-up approach, observing the behaviour of different heads using several documents in our corpus and deriving a general rule. Although these values are apt for our context and they show a remarkable resemblance with other grouping techniques showed in different contexts [10, 11, 19], our method could require a different setting if other BERT models, language or datasets are considered. We are currently studying a method for recognising these thresholds in an automatic way with an unsupervised approach.

3.2. Mean Shift Linker Algorithm

In [10, 12] the connections between pairs of tokens are simply shown with visualization techniques, with lines of different thickness on the basis of the attention weights. However, with relatively long documents with 400 or 500 words the amount of connections increases drastically, making the visualization less understandable and very complex to compute. Therefore, our approach is to directly extract the most important connections among tokens from a specific head. This could help the understanding of the function performed by the head, and simplify

HEAD: 8,8
[CLS] No signs of local recurrence or **lesions**
referable to secondary locali #zations [SEP]

C1 (0.99)	'or'
C2 (0.001)	'[CLS]', 'no', 'signs', 'of', 'local, 'refer', '#able' 'to', 'secondary', 'locali #zations', '[SEP]'

HEAD: 9,8
[CLS] No signs of local recurrence or lesions
refer #able to **secondary** locali #zations [SEP]

C1 (0.32)	'locali'
C2 (0.23)	'to'
C3 (0.08)	'lesions', 'refer', '[SEP]'
C4 (0.05)	'#zations', '#able', 'no'
C5 (0.01)	'signs', 'of', 'local', 'recurrence', 'or', '[CLS]'

Figure 2: Simplified examples of the result of the Mean Shift Linker algorithm for an Offset head (on the left) and for a Local head (on the right)

the visualization process.

Our algorithm for automatically finding these connections is based on Mean Shift [28], a clustering algorithm suitable for density functions and one-dimensional clustering [29]. Given a distribution of attention weights $a_w = [\alpha_1, \alpha_2 \dots \alpha_N]$ for a token w, the different $\alpha_i \in a_w$ are grouped into several clusters depending on their value. In our system, we considered the implementation provided by the standard Machine Learning library Scikit-Learn[1] [30]. Since our clusters are composed by just a few tokens (even just one) and all the tokens needs to be analysed, we set the minimum bin frequency at 1 and the *cluster_all* parameter to *True*. All the other hyperparameters were set to the default values. As we introduced in Section 3.1, one of the most important differences among the head patterns is how the attention weights are distributed across different tokens. For instance, in the Offset heads, a token is connected only to another one. Instead, in the Local heads, the attention can be distributed across several tokens with different degrees of importance. Given that in Mean Shift the number of clusters is defined automatically and it is not selected by the user, this algorithm can easily adapt to such differences. As highlighted in the example showed in Figure 2, on the left we can see how the algorithm selects only two clusters for Offset heads (on the left) and more clusters for the Local heads (on the right). A drawback of this approach is that while the last cluster can be easily discarded as irrelevant and the first considered as important, the role of the intermediate clusters is not immediately understandable. A more detailed analysis of these aspects is going to be conducted as future work.

4. Classification of Radiology Reports

In this section, we describe the real-world context into which we applied BERT, namely the classification of radiology reports. We analyse chest tomography reports, focusing in particular on the possible presence of neoplastic lesions. The potential advantages of a reliable automatic classification of both old and new reports concern diverse areas such as logistics, health care management, monitoring the frequency of follow-up examinations, and collecting cases for research or teaching purposes. The proposed system for report classification is based on a schema defined in strict collaboration with the radiologists [31]. This schema is composed by three levels, that correspond to the main aspects considered by the physicians during the

[1]https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MeanShift.html

	Accuracy	F-Score
Exam Type	96.1	95.9
Result	87.4	84.1
Lesion Nature	73.6	71.6

Table 1
Predictive performance in terms of Accuracy and F-Score. Performance is evaluated in 10-fold cross validation.

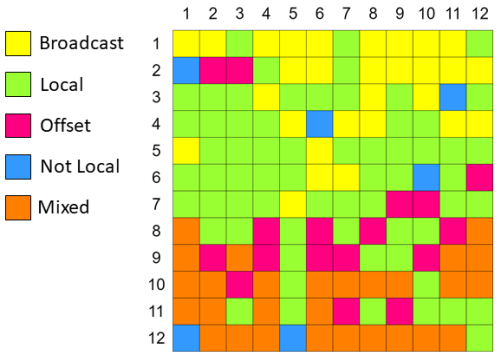
evaluation of a report:

1. **Exam Type** (*First Exam* or *Follow-Up*);
2. **Result** (*Suspect* or *Negative*);
3. **Lesion Nature** (*Neoplastic*, or *lesion with an Uncertain Nature).* This third level is specified only for the Suspect reports.

The dataset is composed by 5,752 classified computed tomography reports. Our reports contain a description (typically without verbs) of what the physicians have seen in the TC images (nodules, lesions, etc.), their relation to previous visits (for instance, if the dimensions are the same with respect to the previous exam) or excluding the presence of specific symptoms or abnormalities. Similarly to other clinical texts, our reports are characterized by a non-standard language, with abbreviations, ungrammatical language, acronyms and typos. This is due to the fact that reports are often written in haste or dictated to a speech recognition software. In addition, abbreviations and acronyms are sometimes idiosyncratic to the specific hospital or department.

In order to see if BERT is effective also in this complex context, we performed the classification task, we adapted the BERT-base Italian model provided by the HuggingFace library[2] by performing the Masked Language Model and Next Sentence Prediction tasks on $10,000$ unclassified reports and next we fine-tune it with our supervised training set. We used Adam as optimizer with learning rate $2 * 10^{-5}$ and batch size 8 for 4 epochs. Performance is evaluated in 10-fold cross validation, therefore training 10 versions of the models using different training and tests and computing the average results. In Table 1, we show the results obtained by our model. For the first two classification levels, the results are higher than 85% in terms of accuracy. For the third level, the performance does not reach the same level of accuracy. This is due mainly to two issues. First, as described in [31], there is no strong agreement among the physicians for identifying *Uncertain Nature* cases and not mistaking them with *Negative* or *Neoplastic* reports. We speculate that these reports contain the most sensible information, therefore their language is the most ambiguous and cryptic. Moreover, it is also probable that some reports can be evaluated as both *Uncertain* or *Negative* depending on the doctor's opinion. Secondly, the third level is only specified for the non negative reports, therefore limiting the number of reports (less than 2000) available for fine-tuning the BERT model. Overall, comparing the results obtained by BERT with the LSTM-based model presented in [31], we can see a small improvement in terms of accuracy and F-Score.

[2]https://huggingface.co/dbmdz/bert-base-italian-cased



(a)

HEAD: 7,4

Invariati millimetrici no #du #li **asp** #ec #i #fici **bilaterali** del massimo diametro di 3mm.

(Millimetric, **asp** #ec #i #fic and **bilateral** no #du #les with maximum diameter of 3mm are unchanged)

HEAD: 9,6

Non segni di **reci** #diva locale nè lesioni parenchimali riferibili a localizzazioni secondarie.

(No signs of local **recur** #rence or lesions referable to secondary localizations)

(b)

Figure 3: (a) Result of our grouping process onto a radiology report. Each row represents an encoding layer of BERT made by 12 heads. (b) Noun-adjective connections (in Italian and with an English adaptation). The token selected is underlined. Its most important connections are in red.

5. Experimental Results

The pre-trained model we used for our classification task has 12 encoding layers and each of which has 12 heads. Thus, we can represent the category of every head in the model as a 12×12 matrix. As mentioned in Section 3.1, our metrics are calculated considering a specific document and therefore some heads (especially in the proximity of the thresholds) may vary their category depending on the document. However, they are a small minority (between 10 and 15) with respect to the total. Analysing Figure 3a we can see some characteristics of the model and our results are quite similar to the ones showed in [19]. First of all, the first 2 encoding layers contain mostly Broadcast heads, which progressively diminish. Instead, the intermediate layers contain mostly Local heads and the last ones are made by a majority of Mixed heads. Offset and Not Local heads are sparse across the model.

In order to select some interesting linguistic characteristics detected by the Italian BERT model in its Local and Not Local heads, we adopted the following approach. First, we calculated the operative tokens, and inspected them searching for interesting patterns. For these tokens, we executed the Mean Shift Linker algorithm, and considered only the first cluster, which contains the highest weights and therefore, it is supposed to highlight the most important relation between pairs of tokens. Finally, on the basis of these results, we formulated an hypothesis on the main linguistic function implemented by the head, and manually annotated some instances of such function. This approach radically differs from the typical probing tasks, where every head is tested with some predefined datasets. In the following sections, we highlight some interesting information that can be extracted from a selection of heads. Please note that we focused our analysis mostly on medical terms, which often are not present in the original vocabulary of the BERT model. Although this does not seem to have a negative impact on our analysis, as future work we will try to perform the same tasks in other domains with a less specific lexicon and evaluate if there are significant differences.

Head	Association	Support	Accuracy
(7,4)	no (nodules) → bilateral	494	78.54%
	portion → ca (caudal)	113	61.06%
(9,6)	local → recur (recurrence)	118	98.30%
	pulmonary → artery	1090	58.07%

Table 2
Results of noun-adjective correlation in two Local heads.

5.1. Adjectives

Considering head $(7, 4)$, i.e., the fourth head of the seventh encoding layer, we have found a significant pattern that connects noun with adjectives. On the top of Figure 3b, we show how the first token of the word *noduli* (nodules, in English) is connected to the first token of the adjective *aspecifici* (aspecific) and the word *bilaterali* (bilateral). Please note that the adjective *millimetrici* (millimetric) is not recognised by the head, which is perhaps due to the fact that the adjective precedes the noun, which is quite uncommon in the Italial language.

On the contrary, head $(9, 6)$ finds connections between adjective and nouns. On the bottom of Figure 3b, we show how the adjective *locale* (local) is connected with the first token of *rediciva* (recurrence).

In order not to limit our analysis only to qualitative examples, we have manually annotated several instances of such correlations and tested how these two heads behave. From our corpus of reports, we annotated the relations between *nodules* and *bilateral*, *portion* and *caudal*, *local* and *recurrence* and *pulmonary* and *artery*. In Table 2, we report the results we obtained. For the relation between *local* and *recurrence*, we have that 98% of the 118 instances annotated were recognised by head $(9, 6)$. For head $(7, 4)$, the relation between *nodules* and *bilateral* has an accuracy of 78.5% across 494 instances. Other relations are recognised with some difficulties, such as the one between *artery* and *pulmonary* or *portion* and *caudal*. These probably suffer from the fact that the words involved are very specific to the medical lexicon, which could not be learned properly due to the limited amount of reports in our dataset.

5.2. Semantic Field

While Local or Mixed heads perform mostly grammatical connections, and therefore concentrate the attention between tokens not so distant from each other, Not Local heads show a completely different behaviour. Given also the fact that they are a minority with respect of the total 144 heads (as can be seen in Figure 3a observing the blue squares), we have inspected them closely. In heads $(2, 1)$, $(3, 11)$ and $(4, 6)$ tokens are mostly connected to themselves or to other occurrences of the same word. This is particularly evident for the word *Non* (not, in English) which appears several times in a report. Usually, each appearance is connected to all the others. There are a few notable exceptions, where very similar tokens are connected, such as *locale* and *locali* (singular and plural versions of the adjective local). Moreover, head $(2, 1)$ is quite noisy and shows random and local connections without any recognisable logic. On the other hand, heads $(12, 1)$ and $(12, 5)$ are very precise, and the vast majority of the tokens points only to themselves, regardless if similar words or other occurrences are present.

Association	Support	Accuracy
lesions → no (nodule)	938	99,98%
texture → paren (parenchyma)	228	92.54%
artery → ##orta (aorta)	606	67.66%
pulmonary → chest	3661	98.55%
inferior → superior	5332	99.83%
segments → portion	131	77.10%
left → right	1784	99.94%

Table 3
Results for the head (6, 10) with examples of relation between words in the same semantic field.

The most interesting phenomenon that we observed regards head (6, 10). Its behaviour can be summarised as follows. If token w is present more than once in the document or there are tokens with very small variations (like singular and plural differences such as *nodule* and *nodules*), then the attention is distributed across all the other occurrences (or little variations) of w. Otherwise, most of the attention is concentrated on synonyms, antonyms or words in the same semantic field of w, like the ones in Table 3. If the two previous conditions are not satisfied, w is connected only to itself.

We investigated further the capability of head (6, 10) to find synonyms or words in the same semantic field. Considering our reports, we analysed the first cluster extracted by the Linker algorithm for each token and discarded when it is connected to itself, or when there are only small variations in terms of characters between the words. In Table 3, we report our results in which we can see some important connections. For instance, the connection between the word *lesions* and the first token of *nodule* (which is a particular kind of lesion) is captured almost every time it appears (99.98% accuracy) over more than 900 instances. In our opinion, the relation between *texture* and the first token of *parenchyma* (with an accuracy of 92.54%) is particularly important, especially given the fact that parenchyma is a very specific word in the anatomy lexicon, describing a particular type of texture. Simple antonyms like *inferior* and *superior* or *left* and *right* are also captured with an high accuracy.

5.3. Negations

We also studied a specifc kind of relations which can be particularly useful in our analysis. In fact, when radiologists evaluate the conditions of a patient, they often write a sentence excluding the presence of something, especially lesions. Sentences like "No focal lesions traceable to secondary locations" are a strong evidence of a negative result of the report, and their individuation could be an important factor in terms of interpretability. Manually inspecting the Local heads, we have identified two of them that can be used to find these patterns: $(9, 8)$ and $(7, 12)$. There is however an important difference: head $(9, 8)$ connects the negative particle with the word which has been denied, while head $(7, 12)$ does the opposite. Nevertheless, instead of a very specific behaviour like the ones showed by [10], when a negation is not present, these heads also can connect adjectives or other tokens, instead of directing the attention into [SEP]. While we cannot straightforwardly claim that these heads are specialised into identifying the negations, we checked if at least they can be used for extracting specific information. Therefore, we annotated the relation between *Non* and *lesioni* (no lesions) and calculated the accuracy of these

heads of recognising; we found that for more than 800 instances, head $(9, 8)$ has an accuracy of 67.9% and $(7, 12)$ reaches 86.0%.

However, while in our context negations are expressed simply with the term "Non" or "nè" (neither), in general negations can be expressed in many different forms, and their detection is a complex task that can require specific models [32]. Moreover, the identification of negations could be limited by the presence of the term *no* not only for introducing a negation but also as the first token of *nodulo* or other similar words. Thus, while the result of head $(7, 12)$ is quite good in this particular case, a more detailed evaluation of the capability of specific heads to detect a negation will be conducted as future work.

6. Conclusions and Future work

We presented an application of fine-tuning the Italian-base BERT model in the context of the classification of radiology reports written in Italian. After verifying its efficacy, we have investigated how its heads behave, and we have proposed to group heads according to their behaviour. An important characteristic of our approach is that it relies only on simple mathematical metrics based on the Jensen-Shannon Divergence, instead of relying on manual observations or clustering. We have also proposed an algorithm based on clustering for automatically extracting the most important connection between words, simplifying the understanding of the characteristics of each head.

Combining these automatic procedures with manual observations, we have found and experimentally evaluated some relevant patterns that can improve the interpretability of BERT for the classification of radiology reports. For instance, in our application it is not sufficient to identify the most important findings or concepts but also to correlate them with their characteristics. For instance, a nodule can be associated with a neoplastic lesion if its margins are irregular or spiculated and not round. Therefore, finding the heads that connect nouns (like margin) and adjectives (like round) could be effectively used for the classification process and its explanation. At the same time, finding that a particular condition is excluded by a negation could be crucial information. Moreover, we have found a head that identifies words in the same semantic field with remarkable accuracy. Although this does not lead to an immediate application for interpretability, this characteristic is another proof of the ability of BERT to capture language properties.

While we have studied and analysed the behaviour of BERT's heads in the radiology context, our techniques are general, and can be easily adapted to other contexts. However, our metrics relies on specific thresholds that could vary on the basis of the document length or its characteristics. As future work, we want to test our techniques more extensively in other applications and using other datasets.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017.

[2] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186.

[3] A. Rogers, O. Kovaleva, A. Rumshisky, A primer in BERTology: What we know about how BERT works, Trans. Assoc. Comput. Linguistics 8 (2020) 842–866.

[4] G. Jawahar, B. Sagot, D. Seddah, What does BERT learn about the structure of language?, in: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 3651–3657.

[5] Y. Goldberg, Assessing BERT's syntactic abilities, CoRR abs/1901.05287 (2019). URL: http://arxiv.org/abs/1901.05287. arXiv:1901.05287.

[6] A. Miaschi, D. Brunato, F. Dell'Orletta, G. Venturi, Linguistic profiling of a neural language model, in: Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020, International Committee on Computational Linguistics, 2020, pp. 745–756.

[7] I. Tenney, D. Das, E. Pavlick, BERT rediscovers the classical NLP pipeline, in: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 4593–4601.

[8] A. Garcia-Silva, J. M. Gomez-Perez, Classifying scientific publications with BERT - is self-attention a feature selection method?, in: Advances in Information Retrieval, Springer International Publishing, Cham, 2021, pp. 161–175.

[9] H. Xu, L. Shu, P. S. Yu, B. Liu, Understanding pre-trained BERT for aspect-based sentiment analysis, in: Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020, International Committee on Computational Linguistics, 2020, pp. 244–250.

[10] K. Clark, U. Khandelwal, O. Levy, C. D. Manning, What does BERT look at? an analysis of BERT's attention, in: Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019, Association for Computational Linguistics, 2019, pp. 276–286.

[11] O. Kovaleva, A. Romanov, A. Rogers, A. Rumshisky, Revealing the dark secrets of BERT, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 4364–4373.

[12] J. Vig, Y. Belinkov, Analyzing the structure of attention in a transformer language model, in: Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019, Association for Computational Linguistics, 2019, pp. 63–76.

[13] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, J. Eisenstein, Explainable prediction of medical codes from clinical text, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018.
[14] Y.-M. Kim, T.-H. Lee, Korean clinical entity recognition from diagnosis text using BERT, BMC Medical Informatics and Decision Making 20 (2020) 1–9.
[15] Y. Si, J. Wang, H. Xu, K. Roberts, Enhancing clinical concept extraction with contextual embeddings, Journal of the American Medical Informatics Association 26 (2019) 1297–1304.
[16] R. Leaman, R. Khare, Z. Lu, Challenges in clinical NLP for automated disorder normalization, Journal of Biomedical Informatics 57 (2015) 28–37.
[17] L. Putelli, A. E. Gerevini, A. Lavelli, R. Maroldi, I. Serina, Attention-based explanation in a deep learning model for classifying radiology reports, in: Artificial Intelligence in Medicine - 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15-18, 2021, Proceedings, volume 12721 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 367–372.
[18] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, E. H. Hovy, Hierarchical Attention Networks for Document Classification, in: NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, The Association for Computational Linguistics, 2016, pp. 1480–1489.
[19] Y. Guan, J. Leng, C. Li, Q. Chen, M. Guo, How far does BERT look at: Distance-based clustering and analysis of BERT's attention, in: Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020, International Committee on Computational Linguistics, 2020, pp. 3853–3860.
[20] A. Miaschi, G. Sarti, D. Brunato, F. Dell'Orletta, G. Venturi, Italian transformers under the linguistic lens, in: Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021, volume 2769 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020.
[21] Z. Meng, F. Liu, E. Shareghi, Y. Su, C. Collins, N. Collier, Rewire-then-probe: A contrastive recipe for probing biomedical knowledge of pre-trained language models, CoRR abs/2110.08173 (2021).
[22] J. Vig, A multiscale visualization of attention in the transformer model, in: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 3: System Demonstrations, Association for Computational Linguistics, 2019, pp. 37–42.
[23] A. Jacovi, Y. Goldberg, Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?, in: Proceedings of the 58th Annual Meeting of the ACL, ACL 2020, 2020, pp. 4198–4205.
[24] S. Jain, B. C. Wallace, Attention is not explanation, in: Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, 2019, pp. 3543–3556.
[25] S. Serrano, N. A. Smith, Is attention interpretable?, in: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 2931–2951.
[26] S. Wiegreffe, Y. Pinter, Attention is not not explanation, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, 2019, pp. 11–20.
[27] B. van Aken, B. Winter, A. Löser, F. A. Gers, How does BERT answer questions? a layer-wise

analysis of transformer representations, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 1823–1832.

[28] D. Comaniciu, P. Meer, Mean shift: A robust approach toward feature space analysis, IEEE Transactions on pattern analysis and machine intelligence 24 (2002) 603–619.

[29] Y. A. Ghassabeh, On the convergence of the mean shift algorithm in the one-dimensional space, CoRR abs/1407.2961 (2014). arXiv:1407.2961.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[31] L. Putelli, A. E. Gerevini, A. Lavelli, M. Olivato, I. Serina, Deep learning for classification of radiology reports with a hierarchical schema, in: Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES-2020, Virtual Event, 16-18 September 2020, volume 176 of *Procedia Computer Science*, Elsevier, 2020, pp. 349–359.

[32] A. Khandelwal, S. Sawant, NegBERT: A transfer learning approach for negation detection and scope resolution, in: Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020, European Language Resources Association, 2020, pp. 5739–5748.