WeNLEX: Weakly Supervised Natural Language Explanations for Multilabel Chest X-ray Classification

Isabel Rio-Torto[*1,2], Jaime S. Cardoso[1,2], Luís F. Teixeira[1,2]

[1]INESC TEC   [2]Universidade do Porto

Abstract

Natural language explanations provide an inherently human-understandable way to explain black-box models, closely reflecting how radiologists convey their diagnoses in textual reports. Most works explicitly supervise the explanation generation process using datasets annotated with explanations. Thus, though plausible, the generated explanations are not faithful to the model's reasoning. In this work, we propose WeNLEX, *a weakly supervised model for the generation of natural language explanations for multilabel chest X-ray classification. Faithfulness is ensured by matching images generated from their corresponding natural language explanations with original images, in the black-box model's feature space. Plausibility is maintained via distribution alignment with a small database of clinician-annotated explanations. We empirically demonstrate, through extensive validation on multiple metrics to assess faithfulness, simulatability, diversity, and plausibility, that* WeNLEX *is able to produce faithful and plausible explanations, using as little as 5 ground-truth explanations per diagnosis. Furthermore,* WeNLEX *can operate in both post-hoc and in-model settings. In the latter, i.e., when the multilabel classifier is trained together with the rest of the network,* WeNLEX *improves the classification AUC of the standalone classifier by 2.21%, thus showing that adding interpretability to the training process can actually increase the downstream task performance. Additionally, simply by changing the database,* WeNLEX *explanations are adaptable to any target audience, and we showcase this flexibility by training a layman version of* WeNLEX, *where explanations are simplified for non-medical users.*

1. Introduction

The importance of explainability, particularly in critical domains like medicine [1, 2] has motivated research into diverse explanation modalities, such as Natural Language Explanations (NLEs). NLEs are textual descriptions that go beyond image captioning, since besides the input image they also need to take into account the underlying model's decision-making process [3–5]. Unlike visual explanations, which are spatially precise but often semantically limited, NLEs are inherently human-understandable, and they are the explanation modality of choice in certain medical contexts where clinicians (e.g., radiologists) convey their findings through textual reports [3, 6, 7].

Despite their promise, current NLE generation approaches face critical limitations. Most rely on fully human-annotated explanation datasets with one NLE per image per diagnosis. This not only imposes a significant annotation burden but also assumes that the reasoning of the model being explained (MBE) is similar to the reasoning of the human annotators. In other words, fully supervised NLEs reflect the annotations, i.e., are plausible, but not the model's true reasoning, i.e., are not faithful. This reliance on explicit supervision also prevents fully explaining pre-trained models, i.e., produce *post-hoc* explanations, as it would require ground-truth NLEs for every combination of correct/incorrect predictions, which differs from model to model. In fact, being able to explain incorrect model decisions is crucial, e.g., for debugging purposes, and it constitutes a clear case where the model reasoning does not align with the human's, hence the definition of an incorrect model prediction.

To address these challenges, we propose *WeNLEX* (pronounced "weenlex"), the first weakly supervised framework for generating NLEs for multilabel chest X-ray classification. *WeNLEX* ensures faithfulness by matching the images generated from the NLEs with the original input images, in the MBE's feature space. Plausibility is maintained through distribution alignment with a small set of clinician-annotated explanations. Through exhaustive validation across multiple dimensions (faithfulness, simulatability, diversity, and plausibility), we show that *WeNLEX*:

1. generates faithful NLEs, requiring as few as five ground-truth explanations per diagnosis

2. can operate in both *post-hoc* and in-model settings, even improving classification performance by 2.21% AUC, demonstrating that interpretability can enhance task per-

---

[*]Correspondence to: `isabel.riotorto@inesctec.pt`

formance

3. offers adaptability to different target audiences, exemplified by a layman version that produces simplified explanations for non-expert users

2. Related Work

The literature on the generation of NLEs for image-only or image-text tasks is vast [8–15]. Hence, there are many architectures for NLE generation, which can be grouped according to different criteria:

- network architecture (the type of neural network): convolutional neural networks [8–10, 16, 17] vs. vision transformers [11–16, 18] for image processing, and recurrent neural networks [8, 9] vs. transformer-based models [11–13, 16–18] for text generation
- modularity (separation between task prediction and explanation modules): modular [8, 9, 11–13, 16, 17] vs. unified [14, 15, 18]
- training paradigm (MBE trained separately or jointly with the explanation model): *post-hoc* [8, 9, 11, 12, 16, 17] vs. in-model [13–15, 18]
- reasoning flow (order between prediction and explanation): *predict-explain* [8, 9, 11, 12, 14–18] vs. *explain-predict* [13]

In the medical domain there are, to the best of our knowledge, only three works on NLE generation [19–21]. Kayser *et al.* [19] introduce the MIMIC-NLE dataset and test different architectures to generate NLEs for multilabel chest X-ray classification, all with the same working principles: i) a vision model extracts features and classifies the input image, and ii) for each predicted diagnosis, the features, the multilabel prediction vector, and the pathology for which the NLE is being generated, are given to a language model to autoregressively generate the NLE.

More recently, Hamza *et al.* [20] proposed a Knowledge Graph Retrieval Augmented Generation (KG-RAG) framework to enhance the LoRA-based [22] finetuning of LLaVA [23]. They build a knowledge graph of relationships between medical entities extracted from the medical reports of MIMIC-CXR [24] using the RadGraph model [25]. For each image, the closest KG entities are retrieved and given to the language module of LLaVa [23], together with the projected image embeddings and the vision classifier predictions. By incorporating domain-specific knowledge, this approach achieves state-of-the-art results on MIMIC-NLE.

All aforementioned works fully supervise the NLE generation process, optimising cross-entropy between generated and human-annotated NLEs. As argued, this leads to plausible but unfaithful explanations that reflect annotations rather than the model's reasoning. Sammani and Deligiannis [5] are the first to tackle this problem in a zero-shot manner. They train a multilayer perceptron (MLP) to map the space of a textual encoder into the vision classifier space. At inference time, learnable prefixes steer an off-the-shelf language model to generate NLEs that maximise the similarity with visual features through the text encoder+MLP. Regardless of being a plug-and-play approach, easily adaptable to any vision classifier, it might not be suitable for problems with less classes (originally the MLP is trained for 1000 classes and it is not clear if it would converge for a smaller number of classes), and it cannot operate in multilabel settings.

Rio-Torto *et al.* [21] propose replacing the commonly used Decoder-only NLE generator with an Encoder-Decoder architecture and showed that it is possible to supervise NLE generation directly in the Encoder latent space. This allows imposing desirable properties on the NLEs via the continuous Encoder latent space, thus avoiding Reinforcement Learning (RL) when teacher forcing is not possible (e.g., when no, or few, ground-truth NLEs are available). In this work, we extend [21] to the weakly supervised setting, thus avoiding the pitfalls of fully supervising the generation of NLEs.

3. Methodology

3.1. MIMIC-NLE Dataset

In the general domain, image captioning datasets abound [26–28]. Datasets with NLEs are more scarce, but do exist, both for text-only [29–32] and vision-language (VL) tasks [8, 10, 12]. Wiegreffe and Marasović [33] provide a comprehensive review on the topic. In the medical domain the same trend emerges: there are some datasets that include medical reports [24, 34–36], but datasets with NLEs are rarer. In fact, MIMIC-NLE [19] is, as far as we are aware, the only dataset with NLEs for chest X-ray classification.

The MIMIC-NLE dataset is automatically extracted from MIMIC-CXR [24] reports using clinically validated rules and the CheXbert labeler [37]. It comprises 38003 image-NLE pairs or 44935 image-label-NLE triplets (one NLE can explain multiple diagnoses/labels). Each image can have up to 10 ($L = 10$) pathologies simultaneously, each with 3 possibilities ($C = 3$): *Positive* (clear evidence of the presence of the pathology), *Uncertain* (the pathology might be present), and *Negative* (clear evidence of the absence of that pathology). These include: i) diagnoses labels, i.e., the labels being explained by the NLEs (*Atelectasis*, *Consolidation*, *Edema*, *Pleural Effusion*, *Pleural Other*, *Pneumonia*, and *Pneumothorax*), and ii) evidence labels, i.e., labels that are part of the evidence that some diagnosis label is present in the input image (*Consolidation*, *Enlarged Cardiomediastinum*, *Lung Lesion*, and *Lung Opacity*). *Consolidation* can be considered both a diagnosis or an evidence label, depending on the interaction with other pathologies (e.g., consolidation as a diagnosis label or consolidation as

evidence for the existence of pneumonia).

A key limitation of this dataset is the strong imbalance in evidence labels, particularly *Lung Opacity* (around 77% of NLEs). Additionally, because evidence extraction relies on the CheXbert labeler [37], around 15% of the NLEs do not have any evidence label; this may be because there are actually no evidence keywords in the original NLE (see examples A and B below) or simply that the evidence present in the NLE is not among the 14 labels that CheXbert is able to predict (see examples C and D below). In such cases evaluating the evidence in generated NLEs is not possible: even incorrect or uninformative outputs (examples A and B) could be deemed valid if CheXbert predicts no evidence labels. To address this, we remove instances without evidence labels. All experiments use this version of the dataset, comprising 36501 NLEs (35600 train, 254 val, 647 test) and 28548 images (27848 train, 198 val, 502 test).

A. Findings most consistent with moderate pulmonary edema.
B. Findings suggesting mild pulmonary vascular congestion.
C. There is again a coarse reticular abnormality favoring the bases and peripheral aspects of the lung, most consistent with pulmonary fibrosis.
D. There is mild vascular congestion consistent with fluid overload.

3.2. Problem Formulation

Let us consider a multiclass multilabel classification task, where the goal is to predict a target $y \in \{1, ..., C\}^L$, where L is the number of labels (of which more than one can be present simultaneously) and C is the number of classes per label, from an input image $x \in \mathbb{R}^{c \times h \times w}$, where c, h, w are the number of channels, height and width, respectively. Let us also consider a classification model, f_{θ_1}, that learns to predict y from x, i.e., $\hat{y} = f_{\theta_1}(x)$. Our goal is to design a system, *WeNLEX*, that generates explanations in natural language, $\hat{e}$, for the predictions of classifier f, such that, $\hat{e} = g_{\theta_2}(x, \hat{y}, f_{\theta_1})$. We adopt the predict-explain paradigm, since we generate one NLE per predicted label and, thus, need to know what the prediction is beforehand.

During training of *WeNLEX* two scenarios are possible: i) f_{θ_1} has been trained and remains frozen (*post-hoc* scenario), or ii) f_{θ_1} is being trained simultaneously with g_{θ_2} (in-model scenario). As will be further detailed in the following sections, the flexibility of *WeNLEX* allows for both.

3.3. WeNLEX

Fig. 1 presents the overall architecture of *WeNLEX*, where for each predicted diagnosis label (either as uncertain or positive), an NLE is generated. *WeNLEX* is inspired by work on unsupervised image captioning [38] and it is built upon our previous work [21], where a pretrained text-only Transformer Encoder-Decoder model is adapted via

Parameter-Efficient Fine-Tuning (PEFT) to receive the output of a multilabel classifier, its features, and the textual description of the diagnosis being explained. In this previous work, under the fully supervised setting (i.e., we had access to one ground-truth NLE per instance), we concluded that using the PEFT method called Multi-Modal LLaMA-Adapter [39] achieved the best results, and that it was possible to supervise NLE generation at the sentence level (i.e., in the latent space of the Encoder), instead of at the word level (i.e., at the output of the Decoder). By doing this, one can impose desirable properties on the generated NLEs through the continuous Encoder latent space, thus avoiding the need for RL. Moreover, training is sped up because word-by-word decoding of the generated sentence embeddings into text is only done during inference. In this work, *WeNLEX*, we extend this previous proposal to the weakly supervised scenario, not only to lower annotation costs, but also to mitigate the previously mentioned lack of faithfulness that arises from fully supervised NLE generation. Thus, *WeNLEX* is designed considering the properties an NLE should have [3, 4, 40, 41]:
- plausibility: sound coherent and logical to a human being
- faithfulness: reflect the model's decision process
- image-relevance: be specific to a given image
- adaptability: adapt to different users

In the following subsections we describe the proposed loss functions that promote each of the aforementioned properties.

3.3.1 Plausibility

In fully supervised NLE generation architectures, plausibility is achieved by approximating generated NLEs with ground-truth NLEs via, e.g., the cross entropy loss. In our case, to be able to do this in a weakly supervised manner (i.e., without a one-to-one correspondence between generated and ground-truth NLEs), we build a database with a fixed number of ground-truth NLEs per diagnosis label. In practice, this database does not have the ground-truth NLEs in textual form, but their embeddings given by the pretrained and now frozen NLE Generator Encoder. Therefore, this becomes a distribution matching problem between the generated and ground-truth NLEs. We experiment with two ways to tackle this: adversarial learning and Maximum Mean Discrepancy (MMD) minimization.

In the first approach, we use a Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP) [42, 43]. WGANP-GP has been proposed to mitigate the mode collapse problems of traditional GANs and is composed of:
1. the generator, g_{θ_2}, generates an NLE for a predicted diagnosis conditioned on the visual features and the prediction of the vision classifier being explained (f_{θ_1}). It
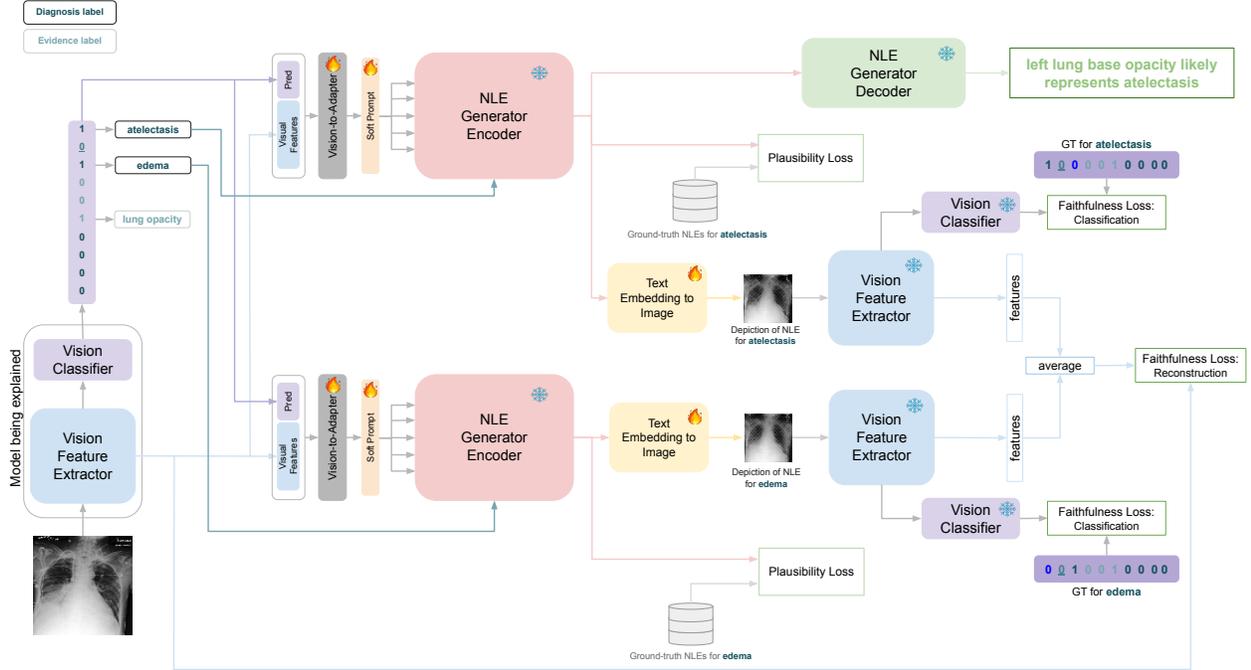
Figure 1. Architecture of *WeNLEX*, a weakly supervised model that generates natural language explanations (NLEs) for a multilabel X-ray classifier. For each predicted diagnosis (e.g., atelectasis, edema), it produces an NLE (only the atelectasis NLE is shown). A pretrained, frozen text-only Encoder–Decoder is adapted with soft prompt tuning (*Vision-to-Adapter* and *Soft Prompt*) to take as input the image features, the entire prediction vector (including diagnosis and evidence labels), and the textual label of the diagnosis being explained. The *NLE Generator Encoder* outputs an NLE embedding, which is compared to ground-truth NLE embeddings for that diagnosis to ensure plausibility (*Plausibility Loss*). Each NLE is also given to a *Text Embedding to Image* model, which generates an image depicting its content. This image is then processed by the model being explained (MBE) to extract features. To enforce faithfulness, the average of these features across all NLEs for an image is compared with the original image features (*Faithfulness Loss: Reconstruction*). Finally, each NLE must recover the MBE's original diagnosis prediction: the MBE's output for the generated image/NLE is compared against the original prediction (*Faithfulness Loss: Classification*). Trainable layers/parameters are represented by the fire icon, while frozen blocks are represented by the snowflake.

learns to minimise the Wasserstein distance between the generated and real NLE embedding distributions by producing NLEs that the critic will score highly as belonging to the real data distribution.

2. the discriminator (or critic), d_{θ_3}, receives the generated NLE embedding and learns to estimate the Wasserstein distance between the real and generated NLE embedding distributions.

Just like in the traditional GAN paradigm, the generator and discriminator play a min-max game, given by:

$$\min_g \max_{d \in \mathcal{D}_{1\text{-Lip}}} \mathbb{E}_{e \sim \mathbb{P}_r}[d(e)] - \mathbb{E}_{\hat{e} \sim \mathbb{P}_g}[d(\hat{e})] \quad (1)$$

where g is the NLE generator, e is the ground-truth NLE embedding, $\hat{e}$ is the generated NLE embedding, d is the discriminator, $\mathcal{D}_{1\text{-Lip}}$ is the set of all discriminators that are 1-Lipschitz continuous, $\mathbb{P}_r$ is the distribution of real NLE embeddings, and $\mathbb{P}_g$ is the distribution of generated NLE embeddings.

The generator loss is given by:

$$\mathcal{L}_g = -\mathbb{E}_{\hat{e} \sim \mathbb{P}_g}[d(\hat{e})] \quad (2)$$

The discriminator loss is given by:

$$\mathcal{L}_d = \mathbb{E}_{\hat{e} \sim \mathbb{P}_g}[d(\hat{e})] - \mathbb{E}_{e \sim \mathbb{P}_r}[d(e)]$$
$$+ \lambda \mathbb{E}_{\tilde{e} \sim \mathbb{P}_{\tilde{e}}}\left[\left(\|\nabla_{\tilde{e}} d(\tilde{e})\|_2 - 1\right)^2\right] \quad (3)$$

$$\tilde{e} = \alpha e + (1 - \alpha)\hat{e}, \quad \alpha \sim \mathcal{U}[0, 1]$$

where $\tilde{e}$ is the interpolation between the ground-truth (e) and the generated ($\hat{e}$) NLE embeddings, $\mathbb{P}_{\tilde{e}}$ denotes the distribution of samples obtained by interpolation, and α is an interpolation factor sampled from a uniform distribution $\mathcal{U}$ on the interval $[0, 1]$. $\nabla_{\tilde{e}} d(\tilde{e})$ is the gradient of the critic used to enforce the 1-Lipschitz constraint. This gradient penalty is controlled by the λ hyperparameter.

Although WGANP-GP works well in theory, in practice it might be difficult to stabilize its training. Moreover, it introduces additional training parameters due to the discriminator. Therefore, we also experiment with MMD minimization.

MMD measures the distance between two distributions through the distance between their embeddings in the reproducing kernel Hilbert space (RKHS):

$$
\begin{aligned}
\widehat{\mathrm{MMD}}^2(X, Y) &= \\
&= \mathbb{E}_{i,j}[k(x_i, x_j)] + \mathbb{E}_{i,j}[k(y_i, y_j)] - 2\,\mathbb{E}_{i,j}[k(x_i, y_j)] \\
&= \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} k(x_i, x_j) + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(y_i, y_j) \\
&\quad - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j) \quad (4)
\end{aligned}
$$

where X is the generated NLE embeddings distribution, Y is the ground-truth NLE embeddings distribution, x_i and x_j (y_i and y_j) are samples from distribution X (Y), m (n) is the number of samples in X (Y), and k is a Gaussian kernel defined by $k(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$. The kernel computes all pairwise distances between X and Y.

In practice, we compute MMD per-label and then average over all labels in the batch so, for each MMD computation, X is composed of the NLEs generated for a given label, while Y consists of the ground-truth NLEs for that label present in the NLE database.

3.3.2 Faithfulness

In addition to being plausible, NLEs should provide enough detail to allow the reconstruction of their corresponding images. Following [38], we do not attempt to reconstruct images in pixel space but instead in the feature space of the MBE. To achieve this, each generated NLE is passed to a module that converts its text embeddings into images. The latter are given to the MBE and the L2 distance between the original and the reconstructed image features is computed. Given the multilabel nature of our primary classification task, there can be several NLEs per image, each referring to a specific pathology. However, the features of the original image contain information on all predicted pathologies (i.e., the features are not disentangled by pathology). So, we actually measure the distance between original image features and the average of all features obtained from each generated image, i.e., from each generated NLE for a given input image (c.f. Fig 1).

Additionally, for an NLE to be faithful it also needs to be able to produce the same decision that gave rise to it in the first place. Although this is somewhat implicitly covered with the feature reconstruction loss detailed above, to

ensure further disentanglement of the NLEs of a given input image, we also compute the classification loss for each generated image. However, two things are important to mention regarding the target of this classification loss: i) it is not the ground-truth of the original image, but the prediction vector of the classifier being explained (as previously explained the NLEs need to be faithful to the model, not to the human annotations), and ii) it is not the full prediction vector of the classifier, but only the prediction vector for that specific diagnosis (the evidence labels are the same for all targets, since there is no way of distinguishing which evidence label led to each diagnosis prediction).

Finally, in the in-model scenario, we ensure that the original classifier is not updated directly when it is used to compute the feature reconstruction and classification losses, i.e., it is updated but only through the backpropagation path through the Text Embedding to Image model and the NLE Generator Encoder. To achieve this, we maintain a frozen copy of the classifier to be used just for the feature extraction needed for the reconstruction and classification losses. Since the original classifier is being updated at every iteration and that could cause training instability, we only update the copy with the newest parameters every 1000 steps.

3.3.3 Image-relevance

Image-relevance is guaranteed in two ways: by giving the visual features extracted by the MBE to the NLE Generator Encoder, and through the feature reconstruction loss, as explained above.

3.3.4 Overall Loss Function

WeNLEX is trained with a combination of the three aforementioned loss functions: i) $\mathcal{L}_{nle\_plaus}$, the plausibility loss (either adversarial or MMD), ii) $\mathcal{L}_{nle\_recons}$, the image feature reconstruction loss, and iii) $\mathcal{L}_{nle\_clf}$, the NLE classification loss. In the in-model case, it also includes $\mathcal{L}_{img\_clf}$, the multiclass multilabel image classification loss, since the classifier is being trained alongside the NLE generation (i.e., in this case *WeNLEX* can be considered a self-explanatory model).

The overall loss function uses the automatic loss weighting method of Cipolla *et al.* [44], thus resulting in the following equations:

$$
\begin{aligned}
\mathcal{L} &= \frac{1}{2\sigma_1^2}\mathcal{L}_{plaus} + \frac{1}{2\sigma_2^2}\mathcal{L}_{nle\_clf} + \frac{1}{2\sigma_3^2}\mathcal{L}_{nle\_recons} \\
&\quad + log(1+\sigma_1^2) + log(1+\sigma_2^2) + log(1+\sigma_3^2) \quad (5)
\end{aligned}
$$

For the in-model case, the overall loss function includes the term $\frac{1}{2\sigma_4^2}\mathcal{L}_{img\_clf} + log(1+\sigma_4^2)$. The σ_i are learnable parameters with an initial value of 1.

4. Experiments

4.1. Implementation Details

Unless otherwise stated, all models are trained with the AdamW optimizer for 50 epochs with a batch size of 16 and a linearly decayed learning rate of 5×10^{-4} with 1000 warmup steps. The final model is chosen based on the lowest validation loss.

4.1.1 Multilabel Classifier

Following previous work on the MIMIC-NLE dataset [19, 21], the vision classifier being explained is the DenseNet-121 [45]. In the *post-hoc* experiments, the classifier is pretrained on the images of the MIMIC-NLE dataset with a class-weighted cross entropy loss, obtaining an AUC of 65.13.

4.1.2 Text AutoEncoder

The text autoencoder (NLE Generator Encoder and NLE Generator Decoder blocks of Fig. 1) is the same as in Rio-Torto *et al*. [21], with the only difference being related to the pretraining dataset. In [21] the encoder was the CXR-BERT model [46], which is a masked language BERT-based transformer model trained on PubMed abstracts [47], clinical notes from MIMIC-III [48] and MIMIC-CXR [24]. Since in this work we are operating under the weakly supervised setting, we assume we do not have access to all MIMIC-NLE sentences. Considering that MIMIC-NLE has been derived from MIMIC-CXR, we wanted to make sure our encoder never had access to MIMIC-CXR. Thus, we take the Pub-MedBERT[1] (BERT model trained from scratch on PubMed abstracts only) [49] and finetune it for 500k steps and an initial learning rate of 2×10^{-5} on sentences of the Interpret-CXR dataset [50] (from which we exclude MIMIC-CXR). Afterwards, we follow the procedure described in [21, 51] and train, still on Interpret-CXR, the single-layer transformer decoder with a denoising auto-encoder objective for 1M steps, batch size of 64, and an initial learning rate of 1×10^{-3}. It achieves a reconstruction BLEU-4 score of 85.9.

4.1.3 Discriminator

When using adversarial learning as the plausibility loss, the whole architecture of *WeNLEX*, except the discriminator, is considered the generator (Vision-to-Adapter, NLE Generator Encoder, Text Embedding to Image model, etc). The discriminator is implemented as a lightweight feed-forward network that receives the NLE and a diagnosis-specific embedding (the latter is obtained by passing each diagnosis

---

[1] https://huggingface.co/microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract

label through the frozen NLE Generator Encoder). The two 768-dimensional vectors are concatenated and passed through a series of fully connected layers with progressively reduced dimensionality (2×768, 512, 256, and 1). The hidden layers are followed by LeakyReLU activations ($\alpha = 0.2$). As is commonly done when training WGANs, for every 5 discriminator updates, the generator is updated once.

4.1.4 Text Embedding to Image Model

The Text Embedding to Image model converts the 768-dimensional NLE embeddings produced by the NLE Generator Encoder to 224×224 grayscale images. The embedding is first projected into a $512 \times 14 \times 14$ feature map using a fully connected layer with ReLU activation. It is then upsampled through four transposed convolutional layers, each doubling spatial resolution while reducing channels ($512 \rightarrow 256 \rightarrow 128 \rightarrow 64 \rightarrow 1$). Batch normalization and ReLU are applied after each intermediate layer, and a final Tanh activation normalizes the output image to $[-1, 1]$.

4.1.5 NLE Database

To build the ground-truth NLE database, we randomly sample n NLEs per diagnosis label, excluding NLEs that explain more than one diagnosis (e.g., "New opacification at the right lung base could be either atelectasis or developing pneumonia."), in order to obtain NLEs as specific to each diagnosis as possible.

4.2. Evaluation Metrics

Evaluating text generation tasks, e.g., image captioning, is notoriously difficult given the inherent variability and ambivalence of natural language, and still somewhat of an unsolved problem in the field [12, 29]. Evaluating NLEs is even harder because, e.g., compared to image captioning, another variable needs to be taken into account: the MBE.

4.2.1 Plausibility

Most works on NLE generation only measure the plausibility of NLEs by comparing them to human-annotated NLEs natural language generation metrics (BLEU [52], METEOR [53], BERTScore [54], etc). Since BERTScore correlates higher with human evaluation [12] and since our NLEs are from the medical domain, we compute their plausibility via BERTScore with the CheXbert [37] model. This score (**CXBS**) is computed only for correctly predicted NLEs (correct diagnosis and evidence labels).

4.2.2 Simulatability

Another way of evaluating NLEs (or other kinds of explanations) is through their utility to an end-user, via simulatability: how well explanations help an observer reproduce the MBE's prediction [55, 56]. This concept appears in the literature under different names and evaluation procedures [13, 15, 57]. In NLX-GPT [15] it is called "Explain-Predict" and the input question and the explanation are given to a language model, which is asked to answer the question (in the context of a visual question-answering task). Wiegreffe *et al.* [57], in the context of a text-only task, measure the additional ability to predict a label that an explanation provides over the input, i.e., the difference between task performance when an explanation is given together with the input vs. when it is not. We adapt these approaches and present the $\hat{y}|NLE$ and $\hat{y}|(img, NLE)$ metrics, which measure whether the NLE alone or together with the image allow a human to reach the same prediction as the MBE. Following previous work [56], we use a proxy for the human observer; we ask the VL foundation model CheXagent [58] to answer if a given label is present in the NLE or in both the NLE and the image. We also compute $\hat{y}|img$ to establish a baseline on the performance of CheXagent on the images alone.

4.2.3 Faithfulness

Most faithfulness metrics target setting with text-only inputs [4, 59, 60]. Both in the realm of visual explanations and NLEs for tasks where the input is text, a popular way to measure faithfulness is by removing parts of the input that the explanation considers important and measuring the decrease in the performance of the MBE [60–62]. Wojciechowski *et al.* [17] adapt this paradigm for the VL domain by asking humans to cover parts of the input image containing the decision rationale indicated in the NLE. We do the same but by leveraging the phrase grounding capabilities of CheXagent: we ask it to ground the NLE in the image and then mask the image in the location(s) given by the bounding boxe(s) produced by CheXagent (see Fig. 2). Following Wojciechowski *et al.* [17], we measure the number of times the multilabel classifier flips its decision when given the masked image (**Flip (%)**) and the absolute difference in the logits of the diagnosis originally predicted ($\mathbf{\Delta_p}$), considering both *Uncertain* and *Positive* logits together.

Following the evaluation protocol of the work that originally proposed the MIMIC-NLE dataset [19], we also compute the CLinical EVidence (**CLEV**) score, which uses the CheXbert [37] model to check if two NLEs refer to the same clinical evidence. However, we introduce two modifications: i) given the big imbalance in the evidence labels of the dataset (see Subsection 3.1), instead of reporting the accuracy, we report the macro-averaged F1-score, and ii) we
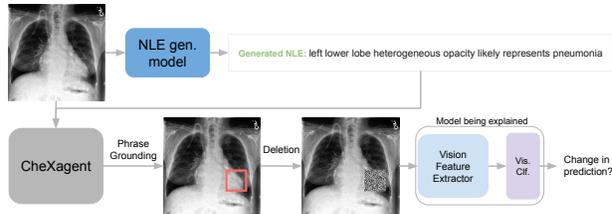

Figure 2. Depiction of the deletion faithfulness metric: an image and a generated NLE are given to CheXagent, which grounds the text in the image. The identified regions are occluded, and the masked image is given to the model being explained (MBE). If the NLE is faithful, occluding these regions should significantly alter the MBE's prediction.

do not compare the evidence of the generated NLE with the evidence of the ground-truth NLE, because the NLE needs to be faithful to the evidence the multilabel classifier predicted, and this might differ from the ground-truth evidence. An example illustrating this can be found in Fig. 3, where the ground-truth refers to the *Lung Opacity* evidence label, but the classifier did not predict this, so the generated NLE cannot and should not mention the presence of a lung opacity. Computing the score in this way has another advantage: since we are now only dependent on the input evidence, the CLEV score can be computed for all generated NLEs, and not only for the correctly predicted ones. In fact, apart from the CXBS, which directly compares against ground-truth NLEs, all other metrics presented throughout the remainder of this work are computed for all generated NLEs.
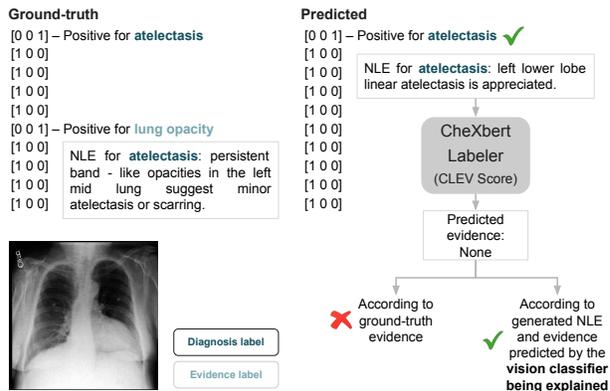

Figure 3. CheXbert identified evidence is correct if it matches the evidence predicted by the model being explained (MBE) (which is also the evidence used to generate the NLE in the first place), but it is incorrect if judged against the ground-truth evidence. Since we want NLEs faithful to the MBE, the target evidence for computing the CLinical EVidence (CLEV) score should be the MBE's predicted evidence.

Table 1. Comparison of *WeNLEX* with the fully supervised baseline of [21] in terms of Faithfulness, Simulatability, Diversity, and Plausibility. ♪ refers to the in-model scenario - the MBE is trained alongside the rest of *WeNLEX*), while ❄ refers to the *post-hoc* scenario (MBE has been pretrained and is frozen during training of *WeNLEX*). "Clf." refers either to the image or the NLE classification loss. "Plaus." refers to the plausibility loss and "Recons." to the feature reconstruction loss. "# NLEs" is the total number of generated NLEs; values in parentheses are for correctly classified cases (diagnosis and evidence). All metrics use the total # NLEs, except CXBS, which uses only the parenthesised values. The ↓ means that for the diversity metrics lower is better. Best results are in bold; second-best are underlined.

Model	Image-Related			NLE-Related											
	Clf.	AUC	$\hat{y}\|img$	Plaus.	Recons.	Clf.	# NLEs	Faithfulness			Simulatability		Diversity ↓		Plausibility
										Deletion	$\hat{y}\|NLE$	$\hat{y}\|(img, NLE)$	S-BLEU	Retrv.	CXBS
								CLEV (F1)	Flip (%)	Δp					
				Fully Supervised											
1	♪	64.08	**83.29**	MSE	✗	✗	1107 (128)	<u>10.37</u>	**97.63**	**0.565**	89.97	88.26	8.785	0.7263	<u>48.00</u>
				Weakly Supervised											
2	❄	65.13	82.47	Adv.	✗	✗	1038 (128)	4.876	94.72	0.2671	71.77	<u>89.11</u>	41.58	0.8079	36.21
3					✗	✗		7.601	91.95	0.3192	86.89	**89.21**	10.36	0.6525	**50.39**
4				MMD	✓	✗		9.303	92.48	0.3236	77.84	86.71	**2.671**	**0.5258**	47.50
5					✓	✓		9.401	91.48	<u>0.3328</u>	<u>90.46</u>	88.82	4.037	<u>0.6286</u>	47.35
6	♪	**67.34**	79.91		✓	✓	1115 (168)	**10.61**	<u>95.65</u>	0.3197	**91.21**	86.91	<u>1.920</u>	0.6704	45.38

4.2.4 Diversity

Finally, we measure the diversity of the generated NLEs. Although two NLEs for the same disease and similar images should be similar, at the same time, they should not be completely equal, since the NLEs also need to be image-relevant. If they were equal, one might assume that having generated the same NLEs is due to correlations and biases in the dataset and not due to the reasoning of the MBE [15]. To measure this, we employ the retrieval-based attack proposed in [15] (**Retrv.**): the 10 most similar images to a query image are retrieved using any feature extraction model (we use MedImageInsight [63]) and *WeNLEX* generates one NLE for each. Then, the cosine distance between these NLEs is computed using Sentence-BERT [64][2]. The lower the distance, the lower the bias, so the better the NLEs will be. To complement this retrieval-based attack, we also compute the per-diagnosis Self-BLEU score (**S-BLEU**), i.e., the BLEU-4 score between all NLEs generated for a given diagnosis. The lower the score, the more diverse the generated NLEs.

4.3. Results

Quantitative results are presented in Table 1. We compare *WeNLEX*, in both *post-hoc* and in-model versions (models 5 and 6, respectively), with the fully supervised model of Rio-Torto *et al.* [21] retrained on our modified version of the MIMIC-NLE dataset (model 1).

We compare the effect of the proposed loss functions under the *post-hoc* setting (models 2-5), since it is the only setting where we are actually comparing the same NLEs (1038 generated NLEs to be precise, of which 128 have the same

diagnosis and evidence labels as their ground-truth counterparts). When comparing the two strategies for promoting plausibility, adversarial learning and MMD minimization (models 2 and 3), it can be seen that MMD minimization increases the CLEV score from 4.876 to 7.601, increases $\hat{y}|NLE$ from 71.77 to 86.89, and increases CXBS from 36.21 to 50.39. Thus, MMD minimization is clearly allowing *WeNLEX* to produce NLEs closer to the ground-truth NLE distribution than the adversarial approach. Moreover, the diversity of the NLEs generated by the adversarial approach is significantly lower than that of the MMD NLEs (41.58 Self-BLEU vs 10.36), which may indicate that the WGAN-GP is collapsing to very similar NLEs for each diagnosis. For these reasons, we chose to continue all remaining experiments with the MMD loss.

Introducing the feature reconstruction loss (model 4) increases the CLEV score from 7.601 to 9.303, the percentage of prediction flips from 91.95 to 92.48, and the absolute difference in the prediction from 0.3192 to 0.3236. This increase in faithfulness comes at the cost of slightly less similarity to ground-truth NLEs (CXBS decreased from 50.39 to 47.50), but it results in a significant increase in the diversity of the NLEs (from 10.36 Self-BLEU to 2.671 and from 0.6525 to 0.5258 in the retrieval metric). Therefore, we conclude that, as hypothesised, introducing the feature reconstruction loss improves the faithfulness of the NLEs. However, both $\hat{y}|NLE$ and $\hat{y}|(img, NLE)$ decrease (from 86.89 to 77.84 and from 89.21 to 86.71, respectively). This might simply be due to the fact that the NLEs of model 4 might be more different from the text that the CheXagent model was trained on, compared to the NLEs of model 3.

Adding the NLE classification loss (model 5) further increases the CLEV score and substantially increases the

$\hat{y}|NLE$ score from 77.84 to 90.46. This comes at a cost of a slight decrease in diversity (e.g., Self-BLEU increases from 2.671 to 4.037), which is to be expected since the classification loss promotes that NLEs for the same disease become similar. However, the increase in CLEV score from 9.303 to 9.401 shows that it brings benefits in terms of ensuring the correct clinical evidence.

In all cases, it can be verified that adding the NLE helps CheXagent predict the label better, as its $\hat{y}|img$ accuracy is always lower than $\hat{y}|(img, NLE)$.

In the in-model scenario (model 6), we obtain the best CLEV, $\hat{y}|NLE$, and Self-BLEU metrics, outperforming the fully supervised scenario, using only 5 ground-truth NLEs per diagnosis. It also achieves the best percentage of flipped decisions amongst the weakly supervised model variants.

Both the classifier trained on its own (models 2-5) and the classifier trained together with the explanation generation part of *WeNLEX* improve upon the AUC of the classifier trained together with the fully supervised pipeline of [21] (65.13/67.34 vs 64.08). When comparing only the *post-hoc* and in-model versions of *WeNLEX* (models 5 and 6), we can see that AUC improves when training the classifier together with the whole model, which shows that, unlike what has been stated many times in the literature [2, 65], adding interpretability does not have to mean a decrease in task performance.

4.4. Ablations

4.4.1 Layers for Feature Reconstruction

We conduct an ablation study on which classifier layer to use for the feature reconstruction loss. In the original Perceptual Loss paper [66], the authors recommend using an intermediate layer for this purpose, since higher layers do not preserve shape as well as lower layers. In our case, there does not seem to clearly exist a layer that is better than the others with regard to the several metrics we evaluate, as can be seen in Table 2. Using higher layers, such as the output of the classification layer or the output of the Global Average Pooling (GAP) layer, leads to less diverse NLEs (Self-BLEU scores of 5.481 and 11.38, respectively). More intermediate layers, such as the output of the 4th and 3rd dense blocks, lead to lower faithfulness scores (CLEV, $\hat{y}|NLE$, $\hat{y}|(img, NLE)$, and Flip. Lower layers like the 2nd and 1st dense blocks improve the CLEV score, while keeping the diversity higher than with the highest level layers (classifier and GAP). They also provide the highest plausibility (CXBS). Thus, we choose to use denseblock2 for all other experiments in this work, since it achieves similar metrics to denseblock1, but with a decreased computational cost (exactly half the number of features).

Table 2. Ablation on the layer used for the feature reconstruction loss. "# Feats" is the dimension of the flattened feature map. "classifier" is the output layer, "gap" the GAP layer, and "denseblock" the outputs of DenseNet layers (with "denseblock1" closest to the input). The underlined layer is used in all other experiments. Best results are in bold; second-best are underlined.

Layer	# Feats.	Faithfulness			Simulatability		Diversity ↓		Plausibility
			Deletion		$\hat{y}|NLE$	$\hat{y}|(img, NLE)$	S-BLEU	Retrv.	CXBS
		CLEV (F1)	Flip (%)	Δp					
classifier	30	**9.623**	91.81	0.3377	91.81	**89.31**	5.481	0.6580	46.96
gap	1024	7.993	**92.39**	0.3279	**92.49**	88.34	11.38	0.6927	46.98
denseblock4	50176	8.887	91.62	**0.3483**	86.13	88.05	2.424	0.5907	46.80
denseblock3	200704	8.631	91.16	0.3344	86.03	87.57	**1.992**	**0.5767**	46.46
denseblock2	401408	9.401	91.48	0.3328	90.46	88.82	4.037	0.6286	47.35
denseblock1	802816	9.463	92.09	0.3385	89.31	88.63	3.654	0.6231	**47.41**

4.4.2 Size of Ground-truth NLE Database

We also ablate the number of clinician-annotated (i.e., ground-truth) NLEs per diagnosis label present in the NLE database. We test *WeNLEX* with 2, 5, 10, and 20 NLEs per diagnosis label. Each experiment is run with 3 different random seeds, and the results (mean and standard deviations) are reported in Table 3.

As expected, using only 2 NLEs per diagnosis label achieves the lowest diversity (Self-BLEU of 8.544 and retrieval score of 0.6852). It also achieves the lowest $\hat{y}|NLE$, $\hat{y}|(img, NLE)$, and Δp. Using 5, 10 or 20 NLEs does not yield significant differences, which attests to *WeNLEX*'s stability and robustness without overreliance on the ground-truth NLEs. Given that using 5 NLEs achieves the highest CLEV score and a lower annotation burden than using more NLEs, all other experiments in this work are performed with 5 NLEs per diagnosis label in the database.

4.5. Audience-Adaptable NLEs

By not needing one NLE per image per diagnosis and simply using a few NLEs per diagnosis, not only does *WeNLEX* significantly lower annotation costs while keeping the generated NLEs faithful, but it also easily allows for the adaptation of the style of the generated NLEs; one simply has to switch the NLEs in the database. This tackles another very important desirable property of NLEs: adaptability to the target audience [6, 67].

To test *WeNLEX*'s ability to generate NLEs for a different target audience (i.e., not radiologists) we ask GPT4-o [68] to convert our NLE database of 5 NLEs per diagnosis label into sentences understandable by lay people, following the prompt used in Zhao *et al.* [69], and then use this converted database to train another version of *WeNLEX*.

The results are presented in Table 4 and Fig. 4. Naturally, the CLEV score decreases, since it uses the CheXbert [37] labeler, which has not seen non-medical sentences during its training. Interestingly, although $\hat{y}|NLE$, $\hat{y}|(img, NLE)$, Flip and Δp all use CheXagent [58], the last two do not seem to be as affected by the layman NLEs as the first two.

Table 3. Ablation on the number of NLEs per diagnosis present in the GT NLE database. The underlined # NLEs is used in all other experiments. Best results are in bold; second-best are underlined.

# NLEs	Faithfulness			Simulatability		Diversity ↓		Plausibility
	CLEV (F1)	Deletion		$\hat{y}\|NLE$	$\hat{y}\|(img, NLE)$	S-BLEU	Retrv.	CXBS
		Flip (%)	Δp					
2	8.663 (0.9852)	**92.76** (0.2796)	0.3530 (0.0342)	85.61 (4.097)	85.64 (0.6675)	8.544 (1.727)	0.6852 (0.0382)	**47.40** (1.672)
5	**9.615** (1.157)	92.14 (0.2282)	0.3539 (0.0324)	91.78 (1.404)	**88.63** (0.4199)	3.560 (1.551)	0.6331 (0.0158)	44.88 (3.371)
10	8.396 (0.1789)	92.33 (0.4732)	**0.3749** (0.0016)	<u>92.29</u> (3.424)	<u>88.34</u> (0.3474)	**1.827** (0.0461)	<u>0.6276</u> (0.0348)	45.22 (1.422)
20	<u>8.672</u> (1.735)	<u>92.52</u> (0.4358)	<u>0.3616</u> (0.0106)	**92.77** (2.569)	88.05 (0.4199)	<u>2.090</u> (0.6986)	**0.5976** (0.0400)	<u>46.45</u> (0.9763)

This might be due to the fact that the location of the findings is given in more similar terms to the original NLEs than the findings themselves (e.g., in the first example of Fig. 4 the location cue "at the left base" has been switched to "of the left lung in the lower field of the lung", while the evidence and diagnosis, "basilar opacity" and "atelectasis or scarring", have been switched to "widened areas" and "barely visible"), such that CheXagent is still able to perform the visual grounding necessary for Flip and Δp, but is no longer able to identify the diagnosis label, which is needed for $\hat{y}|NLE$ and $\hat{y}|(img, NLE)$.

As expected, the diversity is similar in both scenarios, and the plausibility decreases, since the layman NLEs are significantly different from the clinician-annotated NLEs.

Finally, we evaluate the readability of the NLEs (both those generated and those in the database). For this, we use the `textstat` Python package[3] and the `text_standard` consensus metric, which computes the school grade level required to understand a given text, based on several tests (Flesch Reading Ease formula, Flesch-Kincaid Grade Level, Fog Scale, etc). In agreement with the proposed changes, using the simplified NLEs increases readability (the original NLEs are readable at the college level, 13.21, while the simplified NLEs are readable from 10th grade, 10.42). Thus, *WeNLEX* is, in fact, able to adapt to different target audiences.

Table 4. Comparison of *WeNLEX* trained on clinician vs. simplified layperson NLEs. The simplified version shows higher readability (lower required grade level). "Gen." denotes generated NLEs, and "DB" the GT NLE from the medical and non-medical databases.

Database	Faithfulness			Simulatability		Diversity ↓		Plausibility	Readability ↓	
	CLEV (F1)	Deletion		$\hat{y}\|NLE$	$\hat{y}\|(img, NLE)$	S-BLEU	Retrv.	CXBS	Gen.	DB
		Flip (%)	Δp							
Medical	9.401	91.48	0.3328	90.46	88.82	4.037	0.6286	47.35	13.21	15.00
Layman	4.025	91.43	0.2842	64.35	84.30	4.534	0.5295	39.49	10.42	10.46

5. Conclusion

We proposed *WeNLEX*, a flexible weakly supervised framework for generating natural language explanations for mul-
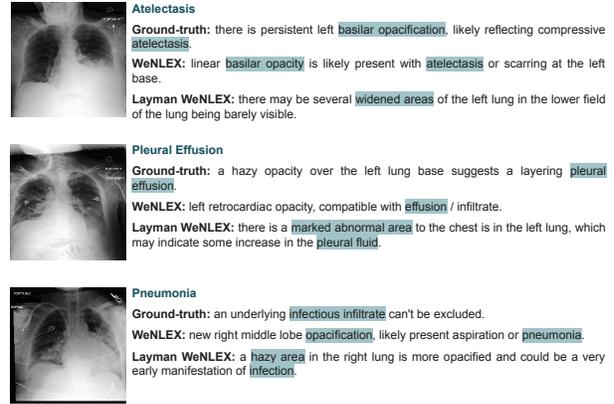

Figure 4. Qualitative examples of *WeNLEX* for three different diagnoses, comparison with ground-truth NLEs, and with the layman version of *WeNLEX*, in which the generated NLEs are simplified to adapt to a non-medical audience.

tilabel chest X-ray classification. Through image feature reconstruction and distribution matching, its explanations are both faithful to the classifier's reasoning and clinically plausible. *WeNLEX*'s versatility allows it to work in a *post-hoc* or an in-model manner. Beyond explainability, the in-model version of *WeNLEX* increases AUC by 2.21%, demonstrating that explainability and task performance can go hand in hand. *WeNLEX* also enables explanations to be tailored to different audiences, from clinicians to lay users, highlighting its broad applicability in real-world medical settings.

References

[1] E. Vayena, A. Blasimme, and I. G. Cohen, "Machine learning in medicine: Addressing ethical challenges," *PLOS Medicine*, vol. 15, pp. 1–4, 2018. 1

[2] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019. 1, 9

[3] I. Rio-Torto, J. S. Cardoso, and L. F. Teixeira, "From Captions to Explanations: A Multimodal Transformer-based Architecture for Natural Language

---
[3] https://github.com/textstat/textstat

Explanation Generation," in *Pattern Recognition and Image Analysis*, 2022, pp. 54–65. 1, 3

[4] C. Agarwal, S. H. Tanneru, and H. Lakkaraju, "Faithfulness vs. Plausibility: On the (Un)Reliability of Explanations from Large Language Models," *arXiv:2402.04614*, 2024. 3, 7

[5] F. Sammani and N. Deligiannis, "Zero-Shot Natural Language Explanations," in *ICLR*, 2025. 1, 2

[6] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019. 1, 9

[7] W. Gale, L. Oakden-Rayner, G. Carneiro, A. P. Bradley, and L. J. Palmer, "Producing radiologist-quality reports for interpretable artificial intelligence," *arXiv:1806.00340*, 2018. 1

[8] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach, "Multimodal Explanations: Justifying Decisions and Pointing to the Evidence," in *CVPR*, 2018. 2

[9] J. Wu and R. Mooney, "Faithful Multimodal Explanation for Visual Question Answering," in *ACL Workshop BlackboxNLP*, 2019, pp. 103–112. 2

[10] V. Do, O.-M. Camburu, Z. Akata, and T. Lukasiewicz, "e-SNLI-VE: Corrected Visual-Textual Entailment with Natural Language Explanations," *arXiv:2004.03744*, 2020. 2

[11] A. Marasović, C. Bhagavatula, J. S. Park, R. Le Bras, N. A. Smith, and Y. Choi, "Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs," in *EMNLP*, 2020, pp. 2810–2829. 2

[12] M. Kayser, O.-M. Camburu, L. Salewski, C. Emde, V. Do, Z. Akata, and T. Lukasiewicz, "e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks," in *ICCV*, 2021, pp. 1244–1254. 2, 6

[13] B. P. Majumder, O. Camburu, T. Lukasiewicz, and J. Mcauley, "Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations," in *ICML*, vol. 162, 2022, pp. 14 786–14 801. 2, 7

[14] B. Plüster, J. Ambsdorf, L. Braach, J. H. Lee, and S. Wermter, "Harnessing the Power of Multi-Task Pretraining for Ground-Truth Level Natural Language Explanations," *arXiv:2212.04231*, 2023. 2

[15] F. Sammani, T. Mukherjee, and N. Deligiannis, "NLX-GPT: A Model for Natural Language Explanations in Vision and Vision-Language Tasks," in *CVPR*, 2022, pp. 8322–8332. 2, 7, 8

[16] Éloi Zablocki, V. Gerard, A. Cardiel, E. Gaussier, M. Cord, and E. Valle, "GIFT: A Framework for Global Interpretable Faithful Textual Explanations of Vision Classifiers," *arXiv:2411.15605*, 2025. 2

[17] A. Wojciechowski, M. Lango, and O. Dusek, "Faithful and Plausible Natural Language Explanations for Image Classification: A Pipeline Approach," in *EMNLP*, 2024, pp. 2340–2351. 2, 7

[18] F. Sammani and N. Deligiannis, "Uni-NLX: Unifying Textual Explanations for Vision and Vision-Language Tasks," in *ICCVW*, 2023, pp. 4636–4641. 2

[19] M. Kayser, C. Emde, O.-M. Camburu, G. Parsons, B. Papiez, and T. Lukasiewicz, "Explaining Chest X-Ray Pathologies in Natural Language," in *MICCAI*, 2022, pp. 701–713. 2, 6, 7

[20] A. Hamza, A. Abdullah, Y. H. Ahn, S. Lee, and S. T. Kim, "LLaVA Needs More Knowledge: Retrieval Augmented Natural Language Generation with Knowledge Graph for Explaining Thoracic Pathologies," *AAAI Conference on Artificial Intelligence*, vol. 39, no. 3, pp. 3311–3319, 2025. 2

[21] I. Rio-Torto, J. S. Cardoso, and L. F. Teixeira, "Parameter-Efficient Generation of Natural Language Explanations for Chest X-ray Classification," in *MIDL*, ser. PMLR, vol. 250, 2024, pp. 1267–1281. 2, 3, 6, 8, 9

[22] Y. Yu, C.-H. H. Yang, J. Kolehmainen, P. G. Shivakumar, Y. Gu, S. R. R. Ren, Q. Luo, A. Gourav, I.-F. Chen, Y.-C. Liu, T. Dinh, A. G. D. Filimonov, S. Ghosh, A. Stolcke, A. Rastow, and I. Bulyko, "Low-Rank Adaptation of Large Language Model Rescoring for Parameter-Efficient Speech Recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2023, pp. 1–8. 2

[23] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *CVPR*, 2024, pp. 26 296–26 306. 2

[24] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, no. 1, p. 317, 2019. 2, 6

[25] S. Jain, A. Agrawal, A. Saporta, S. Truong, D. N. D. N. Duong, T. Bui, P. Chambon, Y. Zhang, M. Lungren, A. Ng, C. Langlotz, P. Rajpurkar, and P. Rajpurkar, "RadGraph: Extracting Clinical Entities and Relations from Radiology Reports," in *NeurIPS Track on Datasets and Benchmarks*, vol. 1, 2021. 2

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *ECCV*, 2014, pp. 740–755. 2

[27] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning," in *ACL*, 2018, pp. 2556–2565.

[28] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts," in *CVPR*, 2021, pp. 3557–3567. 2

[29] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, and P. Blunsom, "e-SNLI: Natural Language Inference with Natural Language Explanations," in *NeurIPS*, vol. 31, 2018. 2, 6

[30] N. F. Rajani, B. McCann, C. Xiong, and R. Socher, "Explain Yourself! Leveraging Language Models for Commonsense Reasoning," in *ACL*, 2019, pp. 4932–4942.

[31] C. Wang, S. Liang, Y. Jin, Y. Wang, X. Zhu, and Y. Zhang, "SemEval-2020 Task 4: Commonsense Validation and Explanation," in *Workshop on Semantic Evaluation*, 2020, pp. 307–321.

[32] S. Aggarwal, D. Mandowara, V. Agrawal, D. Khandelwal, P. Singla, and D. Garg, "Explanations for CommonsenseQA: New Dataset and Models," in *ACL*, 2021, pp. 3050–3065. 2

[33] Wiegreffe, Sarah and Marasović, Ana, "Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing," in *NeurIPS Track on Datasets and Benchmarks*, vol. 1, 2021. 2

[34] P. Chambon, J.-B. Delbrouck, T. Sounack, S.-C. Huang, Z. Chen, M. Varma, S. Q. Truong, C. T. Chuong, and C. P. Langlotz, "CheXpert Plus: Augmenting a Large Chest X-ray Dataset with Text Radiology Reports, Patient Demographics and Additional Image Formats," *arXiv:2405.19538*, 2024. 2

[35] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, "PadChest: A large chest x-ray image dataset with multi-label annotated reports," *Medical Image Analysis*, vol. 66, p. 101797, 2020.

[36] M. D. L. I. Vayá, J. M. Saborit, J. A. Montell, A. Pertusa, A. Bustos, M. Cazorla, J. Galant, X. Barber, D. Orozco-Beltrán, F. García-García *et al.*, "BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients," *arXiv:2006.01174*, 2020. 2

[37] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Ng, and M. Lungren, "Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT," in *EMNLP*, 2020, pp. 1500–1519. 2, 3, 6, 7, 9

[38] Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised Image Captioning," in *CVPR*, 2019. 3, 5

[39] R. Zhang, J. Han, C. Liu, A. Zhou, P. Lu, Y. Qiao, H. Li, and P. Gao, "LLaMA-Adapter: Efficient Fine-tuning of Large Language Models with Zero-initialized Attention," in *ICLR*, 2024. 3

[40] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating Visual Explanations," in *ECCV*, 2016, pp. 3–19. 3

[41] A. Jacovi and Y. Goldberg, "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?" in *ACL*, 2020, pp. 4198–4205. 3

[42] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein Generative Adversarial Networks," in *ICML*, 2017, p. 214–223. 3

[43] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved Training of Wasserstein GANs," in *NeurIPS*, 2017, p. 5769–5779. 3

[44] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics," in *CVPR*, 2018, pp. 7482–7491. 5

[45] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *CVPR*, 2017. 6

[46] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle, H. Poon, and O. Oktay, "Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing," in *ECCV*, 2022, pp. 1–21. 6

[47] U.S. National Library of Medicine, "Pubmed." [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/ 6

[48] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, p. 160035, 2016. 6

[49] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, 2021. 6

[50] J. Xu, Z. Chen, A. Johnston, L. Blankemeier, M. Varma, J. Hom, W. J. Collins, A. Modi, R. Lloyd, B. Hopkins, C. Langlotz, and J.-B. Delbrouck, "Overview of the First Shared Task on Clinical Text Generation: RRG24 and "Discharge Me!"," in *23rd Workshop on Biomedical NLP*, 2024, pp. 85–98. 6

[51] I. Montero, N. Pappas, and N. A. Smith, "Sentence Bottleneck Autoencoders from Transformer Language Models," in *EMNLP*, 2021, pp. 1822–1831. 6

[52] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *ACL*, 2002, p. 311–318. 6

[53] A. Lavie and A. Agarwal, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation

with Human Judgments," in *2nd Workshop on Statistical Machine Translation*, 2007, p. 228–231. 6

[54] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," in *ICLR*, 2020. 6

[55] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," *arXiv:1702.08608*, 2017. 7

[56] P. Hase, S. Zhang, H. Xie, and M. Bansal, "Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language?" in *EMNLP*, 2020, pp. 4351–4367. 7

[57] Wiegreffe, Sarah and Marasović, Ana and Smith, Noah A., "Measuring Association Between Labels and Free-Text Rationales," in *EMNLP*, 2021, pp. 10 266–10 284. 7

[58] Z. Chen, M. Varma, J.-B. Delbrouck, M. Paschali, L. Blankemeier, D. V. Veen, J. M. J. Valanarasu, A. Youssef, J. P. Cohen, E. P. Reis, E. Tsai, A. Johnston, C. Olsen, T. M. Abraham, S. Gatidis, A. S. Chaudhari, and C. Langlotz, "CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation," in *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024. 7, 9

[59] P. Atanasova, O.-M. Camburu, C. Lioma, T. Lukasiewicz, J. G. Simonsen, and I. Augenstein, "Faithfulness Tests for Natural Language Explanations," in *ACL*, 2023, pp. 283–294. 7

[60] N. Siegel, O.-M. Camburu, N. Heess, and M. Perez-Ortiz, "The Probabilities Also Matter: A More Faithful Metric for Faithfulness of Free-Text Explanations in Large Language Models," in *ACL*, 2024, pp. 530–546. 7

[61] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized Input Sampling for Explanation of Black-box Models," in *BMVC*, 2018, p. 151.

[62] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018. 7

[63] N. C. Codella, Y. Jin, S. Jain, Y. Gu, H. H. Lee, A. B. Abacha, A. Santamaria-Pang, W. Guyman, N. Sangani, S. Zhang *et al.*, "MedImageInsight: An Open-Source Embedding Model for General Domain Medical Imaging," *arXiv:2410.06542*, 2024. 8

[64] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *EMNLP*, 2019. 8

[65] P. Atrey, M. P. Brundage, M. Wu, and S. Dutta, "Demystifying the Accuracy-Interpretability Trade-Off: A Case Study of Inferring Ratings from Reviews," *arXiv:2503.07914*, 2025. 9

[66] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," in *ECCV*, 2016, pp. 694–711. 9

[67] Z. Zhu, H. Jiang, J. Yang, S. Nag, C. Zhang, J. Huang, Y. Gao, F. Rudzicz, and B. Yin, "Situated Natural Language Explanations," *arXiv:2308.14115*, 2024. 9

[68] OpenAI, "GPT-4o System Card," *arXiv:2410.21276*, 2024. 9

[69] K. Zhao, C. Xiao, S. Yan, H. Tang, W. K. Cheung, N. A. Moubayed, L. Zhan, and C. Lin, "X-ray Made Simple: Lay Radiology Report Generation and Robust Evaluation," *arXiv:2406.17911*, 2025. 9