LET YOUR HEART SPEAK IN ITS MOTHER TONGUE: MULTILINGUAL CAPTIONING OF CARDIAC SIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cardiac signals convey a significant amount of information about the health status of a patient. Upon recording these signals, cardiologists are expected to manually generate an accompanying report to share with physicians and patients. Generating these reports, however, can be time-consuming and error-prone, while also exhibiting a high degree of intra- and inter-physician variability. To address this, we design a neural, multilingual, cardiac signal captioning framework. In the process, we propose a discriminative multilingual representation learning method, RTLP, which randomly replaces tokens with those from a different language and tasks a network with identifying the language of all tokens. We show that RTLP performs on par with state-of-the-art pre-training methods such as MLM and MARGE, while generating more clinically accurate reports than MLM. We also show that, with RTLP, multilingual fine-tuning can be preferable to its monolingual counterpart, a phenomenon we refer to as the *blessing of multilinguality*.

1 INTRODUCTION

Cardiac signals convey a significant amount of information about the status of a patient. Upon recording these signals, an expert cardiologist is often required to interpret the findings, manually generate an accompanying textual report (Richley & Walters, 2020), and share it with fellow physicians (and patients). Such a report is crucial within medicine, impacting communication between physicians, facilitating clinical decision-making, and holding care teams accountable for their actions (Waegemann et al., 2002). Manually generating these reports, however, can be time-consuming and error-prone, while also exhibiting a high degree of intra and inter-physician variability. Combined, these can detract from a physician's time with a patient, thus decreasing patient satisfaction, and hinder communication between physicians (Hibbard et al., 2001; Keselman & Smith, 2012), thus potentially compromising patient outcomes (Brailer et al., 1997).

Based on these observations, we look to address the following question: *how do we automatically generate clinically accurate reports that reliably summarize cardiac signals?* To address this question, the fields of visual-language representation learning and captioning hold promise. In this setting, rich representations of visual inputs and language are learned to automatically generate captions. Representations of the latter can be learned in a discriminative monolingual setting, as with ELECTRA (Clark et al., 2020), or in a generative multilingual setting, as with MARGE (Lewis et al., 2020). However, a discriminative multilingual framework has yet to be explored. Moreover, previous work has focused exclusively on the captioning of medical images (Hasan et al., 2018; Zeng et al., 2020). It has not explored the captioning of cardiac time-series signals.

In this paper, we address the outlined question by designing a neural, multilingual, cardiac signal captioning system while exploiting a large-scale electrocardiogram (ECG) database. Our contributions are the following: (1) We propose a discriminative multilingual representation learning method, replaced token language prediction (RTLP), in which we randomly replace tokens with those from other languages, and task a network with predicting the language of all tokens. (2) We show that RTLP allows networks to generate reports, in multiple languages, that are more clinically accurate than those generated by masked language modelling (MLM). To the best of our knowledge, we are the first to propose a multilingual cardiac signal captioning system. (3) We conduct extensive studies to show that, with RTLP, multilingual learning outperforms and generates more diverse reports than its monolingual counterpart, a phenomenon we refer to as the *blessing of multilinguality*.

2 RELATED WORK

Language representation learning Representation learning is integral to modern natural language processing. Generative tasks such as MLM have proven effective (Devlin et al., 2019; Liu et al., 2019b), even in the multilingual setting (Conneau & Lample, 2019; Conneau et al., 2020b; Liu et al., 2020). MARGE (Lewis et al., 2020) retrieves documents, in potentially different languages, and attempts to reconstruct a target document. Others propose to jointly learn textual and visual representations (Sun et al., 2019; Lu et al., 2019; Zhang et al., 2020b). Similar to our work is ELECTRA (Clark et al., 2020), where tokens are replaced with those from a generative model, and a network is tasked with discriminating between the original and replaced tokens. This approach, however, is computationally expensive as it requires the learning of both an MLM and discriminative network. Moreover, it does not include the optimization of a multilingual objective. To the best of our knowledge, we are the first to propose a discriminative multilingual representation learning method in the context of cardiac signals.

Multilingual representation learning Pre-training and fine-tuning networks on multiple languages has conferred benefits to NLP tasks (Conneau et al., 2020b; Pratap et al., 2020; Conneau et al., 2020a; Artetxe et al., 2020). For example, Conneau et al. (2020b) and Artetxe et al. (2020) show that multilingual pre-training is more advantageous than its monolingual counterpart in solving downstream NLP tasks. These findings have been partially driven by monolingual datasets that are machine-translated to multiple languages, such as XNLI (Conneau et al., 2018). We similarly translate ground-truth ECG reports into multiple languages in order to guide the generation of multilingual reports. Although Huang et al. (2019a) propose a pre-training setup similar to ours, we explore more languages and define a different pre-training task (RTLP) that is exploited for cardiac signal captioning.

Captioning in healthcare Within biomedical image captioning, most research has focused on chest X-rays (Hasan et al., 2018; Kisilev et al., 2016; Zeng et al., 2020). Liu et al. (2019a) condition their captioning system on the medical topic to be discussed and Wang et al. (2018) propose a multi-level attention model that attends to both the image and the text. More recently, methods such as ClinicalBert (Huang et al., 2019b), BioBert (Lee et al., 2020), and BioELMo (Jin et al., 2019) were shown to benefit biomedical applications (Yoon et al., 2019). Previous work which captioned electroencephalogram (EEG) signals (Biswal et al., 2019; 2020) did not explore a multilingual representation learning method and does not extend to the generation of reports in *multiple* languages. To the best of our knowledge, we are the first to propose the multilingual captioning of cardiac signals.

3 BACKGROUND

3.1 CARDIAC SIGNAL CAPTIONING

We begin by assuming access to a dataset, $\mathcal{D} = \{\boldsymbol{x}_i, \text{cap}_i\}_{i=1}^N$ comprising N cardiac signals, $\boldsymbol{x} \in \mathbb{R}^D$, and their associated captions, $\text{cap} = \{w_s\}_{s=1}^S$, which consist of S words, w_s. The goal of cardiac signal captioning is to generate a caption (report) that reliably summarizes the physiological state of a patient as manifested in a cardiac signal. To extract features from the signal, an encoder, $f_\theta : \boldsymbol{x} \in \mathbb{R}^D \rightarrow \{\boldsymbol{v}_t \in \mathbb{R}^M\}_{t=1}^T$, parameterized by $\boldsymbol{\theta}$, maps a D-dimensional instance, $\boldsymbol{x}$, to a set of T representations, $\{\boldsymbol{v}_t \in \mathbb{R}^M\}_{t=1}^T$, each of which is M-dimensional (see Fig. 1 left).

To convert the captions into a format ingestible by a network, we first convert each word, w_s, in a caption, $\text{cap} = \{w_s\}_{s=1}^S$, to a token, u_s, to form a sequence of tokens $\{u_s\}_{s=1}^S$. Such tokenization can involve lower-casing and stemming words, in addition to removing punctuation. After deriving tokens from all captions in a training set, we form a fixed vocabulary, $V = \{u_i\}_{i=1}^C$, of the C unique tokens where $|V| = C$. We then define an embedding matrix (lookup table), $\boldsymbol{E} \in \mathbb{R}^{C \times M} : u \rightarrow \boldsymbol{e} \in \mathbb{R}^M$, which maps each token, u, to an M-dimensional token embedding, $\boldsymbol{e}$. These embeddings are typically randomly-initialized and learned in an end-to-end manner via gradient descent. As such, every caption can now be represented as a sequence of token embeddings, $\{\boldsymbol{e}_s\}_{s=1}^S$.

To extract features from language, a decoder, $g_\phi : \{\boldsymbol{v}_t\}_{t=1}^T, \{\boldsymbol{e}_s\}_{s=1}^S \rightarrow \{\boldsymbol{h}_s \in \mathbb{R}^M\}_{s=1}^S$, parameterized by $\boldsymbol{\phi}$, attends to all *encoder* representations, $\{\boldsymbol{v}_t\}_{t=1}^T$, while mapping a token em-

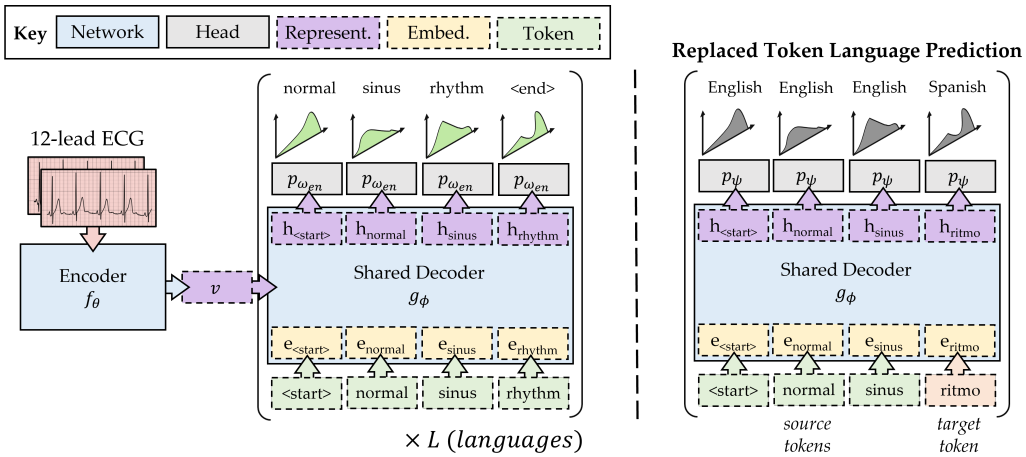

Figure 1: **(Left) Multilingual cardiac signal captioning pipeline.** We feed a 12-lead ECG into an encoder, f_θ, to extract representations, $\boldsymbol{v}$. These are fed, alongside embeddings, $\boldsymbol{e}$, of tokens from a particular language, to a decoder, g_ϕ, to generate token representations, $\boldsymbol{h}$. We feed $\boldsymbol{h}$ into a language-specific head, $p_{\boldsymbol{\omega}_l}$, to generate a caption in a specific language. **(Right) Replaced token language prediction framework**. We randomly replace source tokens with those from a target language and task the network with classifying the language of all tokens. In doing so, we encourage the network to capture relationships between representations of tokens from different languages.

bedding, $\boldsymbol{e}_s$, at step, s, to an M-dimensional token representation, $\boldsymbol{h}_s$. Each token representation, $\boldsymbol{h}_s$, in the sequence of representations, $\{\boldsymbol{h}_s\}_{s=1}^S$, is then fed into a linear classification head, $p_{\boldsymbol{\omega}} : \boldsymbol{h}_s \in \mathbb{R}^M \rightarrow \boldsymbol{y}_s \in \mathbb{R}^C$, to output a probability distribution, $\boldsymbol{y}_s$, over the C tokens in the vocabulary. This forms a sequence of outputs, $\{\boldsymbol{y}_s\}_{s=1}^S$. At each step, s, in the sequence, the goal is to maximize the likelihood of generating the token of the next step, u_{s+1}. Therefore, by identifying the most probable output token at each step, we can form a sentence of words (i.e., a caption).

4 METHODS

4.1 MULTILINGUAL CAPTIONING OF CARDIAC SIGNALS

At the surface, a monolingual captioning framework would appear to be sufficient for physicians communicating in a single language. However, the motivation for a *multilingual* captioning framework, in which reports are simultaneously generated in multiple languages, is threefold. First, a multilingual framework would obviate the cumbersome process of having to train a distinct model for different languages (Conneau et al., 2020a). Second, recent work has demonstrated the benefits of incorporating additional languages into the learning process (Artetxe et al., 2020). Lastly, a multilingual framework exhibits greater flexibility than its monolingual counterpart, as it can always be collapsed, during inference, to generate captions from a single language.

To enable multilingual captioning, we first assume access to L language-specific datasets, $\{\mathcal{D}_l\}_l^L$ where $\mathcal{D}_l = \{\boldsymbol{x}_i, \text{cap}_i^l\}_{i=1}^N$ comprises N cardiac signals, $\boldsymbol{x}$, and captions in a specific language, $l \in \mathbb{L} = \{\text{en}, \text{es}, \cdots\}$, where en and es represent English and Spanish, respectively. Note that the cardiac signals are *shared* across the datasets. To generate captions in multiple languages, we follow the same encoder-decoder approach mentioned in the previous section with one exception. We now replace the single classification head with L language-specific heads to account for the distinct vocabularies of the L languages. In doing so, we exploit recent observations that demonstrated the utility of having a network with parameters that are both language-specific and shared across languages (Zhang et al., 2020a).

Formally, for each language, l, we have a linear classification head, $p_{\boldsymbol{\omega}_l} : \boldsymbol{h}_s^l \in \mathbb{R}^M \rightarrow \boldsymbol{y}_s^l \in \mathbb{R}^C$, with $\boldsymbol{\omega}_l \in \{\boldsymbol{\omega}_{\text{en}}, \boldsymbol{\omega}_{\text{es}}, \cdots\}$ reflecting language-specific parameters. Each head maps the token representation, $\boldsymbol{h}_s^l$, at each step, s, in the sequence to a probability distribution, $\boldsymbol{y}_s^l$, over C tokens

where $C \in \{|V_{\text{en}|}, |V_{\text{es}}|, \cdots\}$ reflects the size of a language-specific token vocabulary. When doing this for each step in the sequence, we arrive at a set of probability distributions, $\{\boldsymbol{y}_s^l\}_{s=1}^S$. As with traditional language models, at each step in the sequence, $s \in [1, S]$, we maximize the likelihood of observing the next token in the sequence, u_{s+1}^l, from a particular language, $l \in \mathbb{L}$. Therefore, for a mini-batch of B captions in a single language, we would optimize the categorical cross-entropy loss at each step, s. To extend this to L languages, we load L mini-batches and optimize the following *multi-task* categorical cross-entropy loss.

$$\mathcal{L}_{\text{multilingual}} = -\frac{1}{LBS} \sum_{l \in \mathbb{L}}^{L} \sum_{i=1}^{B} \sum_{s=1}^{S} \log p_{\boldsymbol{\omega}_l}(y_{i,s}^l = u_{i,s+1}^l) \tag{1}$$

4.2 Replaced Token Language Prediction

To facilitate achieving the downstream task of multilingual cardiac signal captioning, we design a discriminative multilingual pre-training task that learns the decoder parameters, ϕ, and the token embeddings, $\boldsymbol{e}$. At a high-level, this task involves randomly selecting tokens in a sequence, replacing them with semantically-similar tokens from a different language, and tasking a network with classifying the language of all tokens (see Fig. 1 right).

The intuition behind our framework is that a network exposed to semantically-similar tokens from distinct languages which share the same context (neighbouring tokens) can learn that such tokens are indeed similar to one another. We hypothesize that encouraging this behaviour can lead to the learning of token representations that are beneficial for the downstream task of *multilingual* captioning since these tokens (from different languages) are likely to arise in the same generated report, albeit in a different language. As such, the generated multilingual reports might be more likely to contain plausible tokens, and thus be clinically accurate. From hereon forward, we refer to this method as replaced token language prediction (RTLP) and describe its mechanics next.

Source token selection Given tokens, $\{u_s^{l_{src}}\}_{s=1}^S$, in a sequence of length, S, from a source language, $l_{src} \in \mathbb{L}$, we first sample K distinct steps, $\{s_k\}_{k=1}^K$ where $s_k \in [1, S]$ from a uniform distribution, $\mathcal{U}$. We then replace the corresponding source tokens, $\{u_{s_k}^{l_{src}}\}_{k=1}^K$, with those from a target language, $\{u_{s_k}^{l_{tgt}}\}_{k=1}^K$. We outline how to select the target language and token next.

Target language and token selection For each source token, $u_{s_k}^{l_{src}}$, we sample a target language, $l_{tgt} \sim \mathcal{U}(\mathbb{L}')$, uniformly at random from the set of remaining languages $\mathbb{L}' = \mathbb{L} \setminus l_{src}$ where $|\mathbb{L}'| = L - 1$. Given l_{tgt}, we now sample a target token, $u_{s_k}^{l_{tgt}} \sim \mathcal{U}(V_{l_{tgt}})$, uniformly at random from the language-specific vocabulary of tokens, $V_{l_{tgt}}$. Such random sampling, however, can lead to the selection of a target token that is semantically different from the source token. As such, the network may discriminate between source and target tokens by using a detrimental shortcut that is based on semantics instead of language.

To avoid this behaviour, we instead adopt a strategy where the target token is likely to be semantically similar to (e.g., a noisy translation of) the source token. Formally, we quantify the cosine similarity, sim_j, between the source token embedding, $\boldsymbol{e}_{s_k}^{l_{src}} \in \mathbb{R}^M$, and the embedding, $\boldsymbol{e}_j^{l_{tgt}} \in \mathbb{R}^M$ of each token, $u_j^{l_{tgt}} \in V_{l_{tgt}}$ in the vocabulary of the target language. By taking the softmax of these similarities, we form a categorical distribution, q, with elements, q_j, from which we sample $u_{s_k}^{l_{tgt}}$. As the token embeddings become more meaningful during training, the sampled target token is more likely to be semantically similar to the source token.

$$u_{s_k}^{l_{tgt}} \sim q \quad , \quad \text{where} \quad q_j = \frac{\exp^{(\text{sim}_j)}}{\sum_m^{|V_{l_{tgt}}|} \exp^{(\text{sim}_m)}} \quad , \quad \text{sim}_j = \frac{\boldsymbol{e}_{s_k}^{l_{src}} \cdot \boldsymbol{e}_j^{l_{tgt}}}{|\boldsymbol{e}_{s_k}^{l_{src}}||\boldsymbol{e}_j^{l_{tgt}}|} \tag{2}$$

Objective function Equipped with a sequence of tokens in source and target languages, we predict the language of such tokens. To do so, we define a classification head, $p_{\boldsymbol{\psi}} : \boldsymbol{h} \rightarrow \boldsymbol{y} \in \mathbb{R}^L$, parameterized by $\boldsymbol{\psi}$, that maps each token representation, $\boldsymbol{h}$, to a probability distribution, $\boldsymbol{y}$, over L languages. Formally, given a mini-batch of B captions of length, S, in the source language, l, we

optimize the categorical cross-entropy loss. This is achieved for the tokens in the source ($r = 0$) and target ($r = 1$) languages. To extend this to L source languages, we load L mini-batches and optimize the following *multi-task* categorical cross-entropy loss where $\mathbb{1}$ is the indicator function.

$$\mathcal{L}_{\text{RTLP}} = -\frac{1}{LBS}\sum_{l\in\mathbb{L}}^{L}\sum_{i=1}^{B}\sum_{s=1}^{S}\mathbb{1}_{r=0}\cdot\log p_\psi(y_{i,s}=l_{src}|u_{i,s}^{l_{src}})+\mathbb{1}_{r=1}\cdot\log p_\psi(y_{i,s}=l_{tgt}|u_{i,s}^{l_{tgt}}) \quad (3)$$

5 Experimental Design

Datasets We evaluate our framework on the PTB-XL dataset (Wagner et al., 2020), the *only* publicly-available dataset which comprises ECG signals with a corresponding clinical report. Each ECG signal (from a total of 18,885 patients) is also associated with cardiac arrhythmia labels which we group into 5 classes (Strodthoff et al., 2020). We provide further details in Appendix A.1.

Representation learning of cardiac signals Supervised pre-training continues to be an effective way to learn rich, generalizable representations. To that end, we pre-train the *encoder*, f_θ, to map 12-lead ECG signals to abnormalities in the functioning of the heart (cardiac arrhythmias). This choice is motivated by our observation that ECG reports are likely to reflect the pathology of a cardiac arrhythmia. Therefore, learning representations of cardiac signals that can discriminate between cardiac arrhythmias may be beneficial for the downstream task of cardiac signal captioning.

Representation learning of clinical reports To implement RTLP, we exploit L language-specific tokenizers from SpaCy to tokenize the ECG reports. For simplicity, we lower-case the text and remove any punctuation. By keeping track of unique tokens, we form L distinct vocabularies. Each vocabulary also includes language-specific tokens to indicate the start and end of the report, and the $[\text{PAD}]$ and $[\text{OOV}]$ tokens to refer to padded entries and tokens observed during inference that are not seen during training, respectively. We also introduce the $[\text{MASK}]$ token where appropriate.

Multilingual cardiac signal captioning After pre-training the encoder, f_θ, and the decoder, g_ϕ, independently of one another, we exploit the learned parameters, $\{\theta, \phi\}$, and token embeddings to solve the task of cardiac signal captioning (see Fig. 1 left). After experimenting with several variants of our framework, we chose to *freeze* the encoder parameters and extract multiple representations per cardiac signal. These encoder representations are used in the cross-attention mechanism of the Transformer decoder. Multilingual captioning also requires a ground-truth ECG report to be available in multiple languages. Since such *paired* reports do not exist, we translate the original set of reports in English (en) to six languages $\{\text{German (de)}, \text{Greek (el)}, \text{Spanish (es)}, \text{French (fr)}, \text{Italian (it)}, \text{Portuguese (pt)}\}$ using the Google Translate API, a strategy similar to that adopted by Conneau et al. (2018). We open-source these reports[1] and provide further details in Appendix B. Although translated reports can be imperfect ground-truth reports, we hypothesize (and indeed show) that the net effect of multilinguality on downstream performance is advantageous.

Baseline Methods As we are the first to propose the task of multilingual captioning of cardiac signals, previous methods do not exist for comparison. However, we compare our pre-training method, RTLP, to the following: 1) **MLM** (Devlin et al., 2019), a masked language modelling pre-training objective where the decoder is tasked with identifying masked tokens, 2) **ELECTRA** (Clark et al., 2020), a pre-training objective where the decoder is tasked with identifying whether tokens have been replaced with those from an MLM model, and 3) **MARGE** (Lewis et al., 2020), a multilingual generative language representation learning approach where source documents in various languages are retrieved to generate a similar yet distinct target document (see Appendix C.3 for more details on how we adapted these methods).

Evaluation Metrics We quantify the quality of the generated reports by comparing the degree of overlap of tokens in such reports to those in the ground-truth report. Specifically, we use the $\text{BLEU}-1$ (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), and $\text{ROUGE}-\text{L}$ (Lin, 2004) scores (see Appendix C.4 for more details).

[1] Code and data can be found at: `https://tinyurl.com/CardiacSignalCaptioning`

Hyperparameters We conduct our experiments using PyTorch (Paszke et al., 2019). We pre-train the encoder and decoder with a patience value of 10 and 25 epochs, respectively, on the validation loss. When transferring the parameters to the task of cardiac signal captioning, we use those associated with the lowest validation loss. When fine-tuning, we checkpoint the parameters associated with the highest validation BLEU score. We implement a greedy decoding mechanism by taking the argmax of the probability distribution over tokens at each step. We leave for future work other decoding mechanisms such as beam search (Graves, 2012).

6 EXPERIMENTAL RESULTS

6.1 QUANTITATIVE EVALUATION OF GENERATED REPORTS

We begin by quantitatively evaluating the ability of the pre-training methods to generate high quality multilingual reports. In Table 1, we present the $\text{BLEU}-1$, METEOR, and $\text{ROUGE}-\text{L}$ scores of reports generated in seven different languages.

We find that RTLP performs on par with state-of-the-art generative pre-training methods. For example, on average, RTLP achieves $\text{BLEU}-1 = 28.5$ whereas MLM and MARGE achieve $\text{BLEU}-1 = 29.4$ and 28.9, respectively. This finding holds across languages and evaluation metrics. Furthermore, we find that RTLP outperforms the state-of-the-art *discriminative* pre-training method, ELECTRA. For example, on average, RTLP and ELECTRA achieve $\text{ROUGE}-\text{L} = 33.4$ and 0.5, respectively. We note that we were unable to achieve satisfactory performance with ELECTRA despite extensive experimentation (please see GitHub for reproducibility). One hypothesis for this stems from the difficulty of optimizing its objective function which comprises a generative *and* discriminative term. We also find that, regardless of the pre-training method implemented, performance varies significantly across languages. For example, RTLP achieves $\text{BLEU}-1 = 19.8$ and 33.5 on Greek (el) and English (en) reports, respectively. We hypothesize that this is due to a high level of dissimilarity between the Greek vocabulary and that of the remaining languages. This, in turn, reduces the amount of knowledge transferred from the other languages to the Greek language. Overall, our findings suggest that RTLP can be an effective discriminative multilingual pre-training method. Qualitative evidence to support this claim is provided in the next section.

Table 1: **Multilingual cardiac signal captioning performance of pre-training methods.** Results are shown on the test set across five seeds. The standard deviation is shown in brackets. For clarity, we have highlighted our method in gray. We find that RTLP performs on par with state-of-the-art language pre-training methods, MLM and MARGE.

Language Pre-training Method	German (**de**)	Greek (**el**)	English (**en**)	Spanish (**es**)	French (**fr**)	Italian (**it**)	Portuguese (**pt**)	**Average**
BLEU-1								
MLM (Devlin et al., 2019)	25.9 (0.6)	20.5 (0.3)	31.3 (0.5)	33.2 (0.8)	29.7 (0.6)	30.3 (0.2)	34.9 (0.7)	29.4 (4.6)
ELECTRA (Clark et al., 2020)	0.1 (0.1)	0.2	0.2 (0.2)	0.3 (0.1)	0.6 (0.1)	0.5 (0.1)	0.5 (0.1)	0.3 (0.2)
MARGE (Lewis et al., 2020)	24.9 (1.0)	19.5 (0.9)	30.8 (0.5)	32.9 (0.5)	29.7 (0.6)	29.4 (0.5)	34.5 (1.0)	28.9 (4.8)
RTLP	25.4 (1.1)	19.8 (0.6)	30.0 (0.7)	33.1 (0.9)	28.3 (0.8)	30.0 (0.1)	33.5 (1.0)	28.5 (4.5)
METEOR								
MLM (Devlin et al., 2019)	36.7 (1.0)	23.6 (0.2)	37.3 (1.1)	38.6 (0.6)	33.5 (0.7)	33.9 (0.7)	38.8 (0.7)	34.6 (5.0)
ELECTRA (Clark et al., 2020)	0.3 (0.5)	0.2 (0.1)	0.2 (0.1)	0.5 (0.4)	1.1 (0.3)	0.9 (0.2)	0.5 (0.2)	0.5 (0.4)
MARGE (Lewis et al., 2020)	35.6 (1.5)	22.2 (1.0)	36.5 (0.8)	37.1 (0.6)	33.1 (1.0)	32.9 (0.9)	37.8 (1.1)	33.6 (5.1)
RTLP	36.5 (0.7)	22.6 (1.0)	36.0 (0.9)	38.5 (0.8)	32.4 (0.4)	33.7 (0.9)	37.6 (0.6)	33.9 (5.1)
ROUGE-L								
MLM (Devlin et al., 2019)	34.6 (0.8)	11.4 (1.9)	28.5 (0.2)	39.3 (1.3)	34.5 (0.8)	36.9 (0.5)	39.1 (0.6)	33.5 (9.3)
ELECTRA (Clark et al., 2020)	0.2 (0.3)	0	0.2	0.5 (0.3)	1.0 (0.2)	0.8 (0.1)	0.5 (0.1)	0.5 (0.4)
MARGE (Lewis et al., 2020)	33.2 (0.7)	11.1 (2.3)	38.1 (0.5)	39.2 (0.6)	34.4 (0.8)	36.1 (0.6)	39.0 (0.5)	33.0 (9.4)
RTLP	34.0 (1.1)	11.6 (2.3)	36.3 (0.9)	39.1 (1.2)	33.1 (0.9)	36.5 (1.0)	37.3 (0.9)	32.6 (8.9)

6.2 QUALITATIVE EVALUATION OF GENERATED REPORTS

We also manually inspect the multilingual clinical reports generated by RTLP. Ideally, such reports should accurately reflect clinical information. In Fig 2, we present a 12-lead ECG segment alongside the multilingual ground-truth reports and those generated by RTLP and MLM.

There are three main takeaways from Fig 2. First, RTLP allows networks to generate reports that accurately capture the aberrant morphology (shape) of the ECG signal. For example, in Spanish (es), the ground-truth and RTLP-generated reports both explicitly mention the cardiac abnormality *"bloqueo de rama izquierda"* (left bundle branch block). Second, we find that RTLP-generated reports manage to capture critical aspects of the ECG signal that their MLM counterparts struggle with. For example, in English (en), both the ground-truth and RTLP-generated reports both explicitly mention *"left hypertrophy possible"*, whereas this phrase is noticeably absent from the MLM-generated report. Such an absence is problematic as it might result in a physician overlooking this aspect, and thus failing to act accordingly. Lastly, we find that MLM-generated reports can include additional *erroneous* phrases which are neither present in the ground-truth report nor in the RTLP-generated report. For example, in Portuguese (pt), the MLM-generated report mentions *"atrial flutter"*, which is noticeably (and correctly) absent from the remaining reports. Similar erroneous inclusions can also be found in the German (de) and French (fr) reports. This is also problematic since physicians can be misled by such statements, potentially resulting in unnecessary medical treatments. These findings indicate that RTLP allows networks to generate reliable and clinically accurate reports.

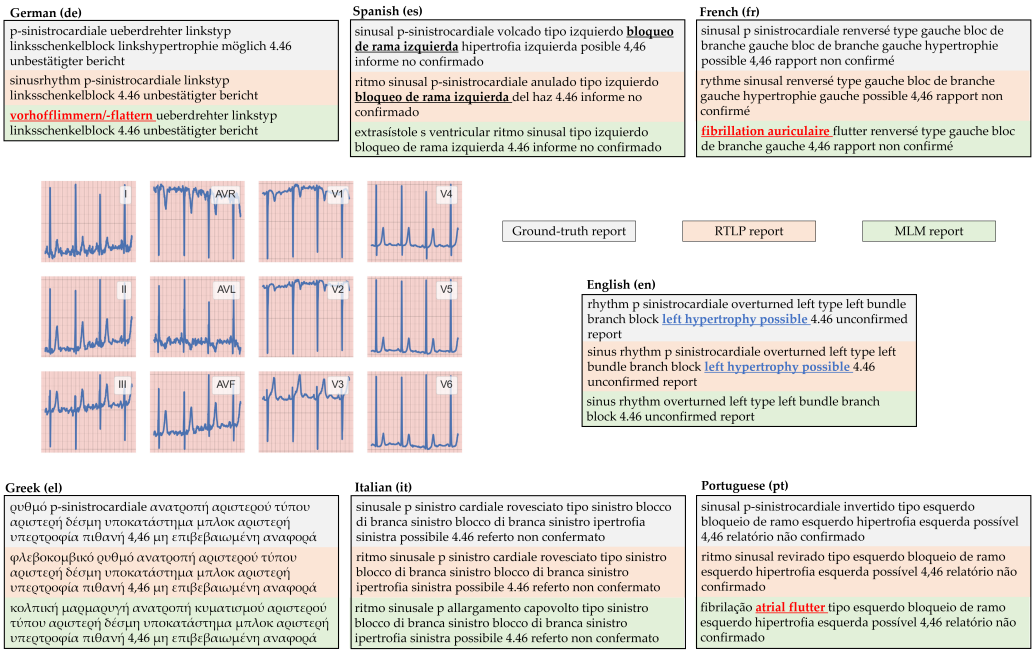

Figure 2: **12-lead ECG, multilingual ground-truth reports, and those generated by RTLP and MLM.** We show some phrases in **bold** which exhibit a high level of agreement in both the ground-truth report and that generated by RTLP, in blue which are captured by RTLP and *not* captured by MLM (false negatives), and in red which MLM erroneously includes (false positives). Overall, we show that RTLP can generate reports that accurately capture the pathology of the cardiac signal.

6.3 Quantification of the Diversity of Generated Reports

So far, we have shown that RTLP is capable of generating clinically accurate and plausible reports. It could be argued that generated reports should also exhibit a certain level of diversity in text (Li & Jurafsky, 2016). To illustrate why, consider a constrained system that simply regurgitates the same caption regardless of the input cardiac signal. Such a system would be of minimal clinical value due to its inability to distinguish between the nuances of the cardiac signals. To quantify the diversity of generated reports, we exploit the Self-BLEU metric (Zhu et al., 2018), which measures the BLEU score between all pairs of generated reports. Intuitively, the lower the value of this metric (↓ Self-BLEU), the higher the diversity of the reports (↑ diversity). In Fig. 3, we present the Self-BLEU score for three different pre-training methods across all languages in the multilingual setting.

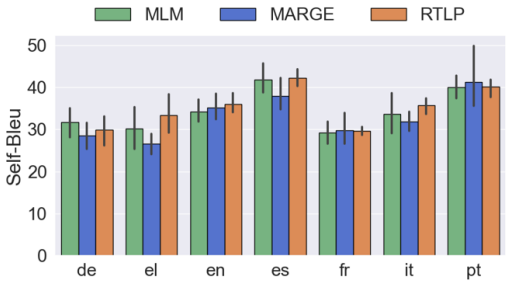

Figure 3: **Self-BLEU (diversity) of the reports generated by various pre-training methods.** ↑ Self-BLEU implies increased similarity between reports and thus ↓ diversity. Error bars indicate the standard deviation across five seeds. We show that the diversity of generated reports differ significantly across languages and that the reports generated by different methods exhibit a similar level of diversity, irrespective of the language.

There are several takeaways from Fig. 3. First, the degree of diversity exhibited by the generated reports differs across the languages. For example, the French (fr) reports, on average, exhibit the lowest Self-BLUE ($\approx$ 30) and thus highest diversity whereas the Spanish (es) reports exhibit the least diversity (Self-BLEU $\approx$ 40). We hypothesize that this could be due to the distinct vocabulary sets of these respective languages. Second, we find that all three methods, MLM, MARGE, and RTLP, generate reports with a similar level of diversity. For example, when generating English (en) reports, these methods achieve Self-BLEU $\approx$ 35.

6.4 Investigation of the Curse of Multilinguality

Multilingual neural systems can experience the *curse of multilinguality* (Conneau et al., 2020b). Concisely, this attributes the potentially poorer performance of multilingual models relative to their monolingual counterparts to interference between the various languages. Intuitively, tasking a network with generating reports in multiple languages, analogous to multi-task learning (Caruana, 1993), can be too demanding and thus hinder its ability to generate sensible reports. We explore this curse from the lens of the clinical accuracy and diversity of the generated reports. To do so, we first pre-train our networks, as per usual, and fine-tune them in the *monolingual* setting. We then compare the performance of networks and the diversity of the generated reports in the multilingual setting to those in the monolingual setting. In Fig. 4, we present such a comparison, illustrating $\text{BLEU}-1$ (top row) and Self-BLEU (bottom row) of reports generated by MARGE and RTLP.

In Fig. 4 (top row), we find that MARGE does not experience the curse of multilinguality when evaluated along the dimension of performance. This can be seen by the similar performance achieved by MARGE irrespective of whether it is fine-tuned in the monolingual or multilingual setting. For example, when generating Spanish (es) reports, MARGE achieves $\text{BLEU}-1 \approx 32$ in both settings. Such a finding suggests that incorporating additional languages into the fine-tuning process has little to no effect on the quality of the generated reports. In contrast, we find that RTLP benefits significantly from the inclusion of multiple languages. This can be seen by the higher generalization performance ($\uparrow$ $\text{BLEU}-1$) achieved by RTLP in the multilingual setting than in the monolingual setting. For example, for Spanish (es) reports, the network in both settings achieves $\text{BLEU}-1 \approx 32$ and $\approx$ 18, respectively. This reflects almost a two-fold improvement. We denote this marginal benefit of multilinguality as the *blessing of multilinguality*. We hypothesize that such benefits stem from the transfer and sharing of knowledge across languages. We also note the poorer performance of RTLP in the monolingual setting relative to the multilingual setting. For English (en) reports, RTLP achieves $\text{BLEU}-1 \approx 13$ and ≈ 31, respectively. We hypothesize that this is due to the relative degree of importance placed by the model on particular languages during multilingual pre-training with RTLP. In other words, pre-training may implicitly weight languages differently. As such, the learned language-specific token representations may differ in their expressiveness. In light of this, networks fine-tuned in the monolingual setting may perform poorly.

In Fig. 4 (bottom row), we find that, on average, MARGE generates reports with a similar level of diversity irrespective of whether it is fine-tuned in the monolingual or multilingual setting. For

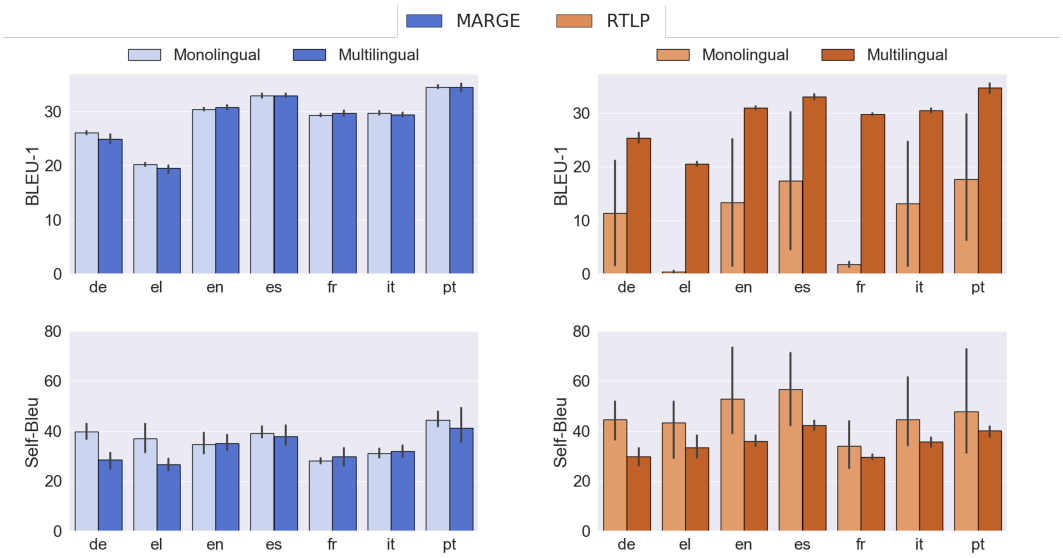

Figure 4: **Effect of multilinguality on the (top row) quality and (bottom row) diversity of the generated reports.** The multilinguality setting involves simultaneously generating reports in all seven languages, $\mathbb{L} = \{\text{de}, \text{el}, \text{en}, \text{es}, \text{fr}, \text{it}, \text{pt}\}$. Results are shown for all reports and across five random seeds. We show that, through the lens of report quality and diversity, MARGE does not consistently experience the curse of multilinguality. In contrast, RTLP benefits significantly from incorporating multiple languages into the fine-tuning process, a phenomenon denoted by the *blessing of multilinguality*.

example, although MARGE generates more diverse ($\downarrow$ Self-BLEU) German (de) reports in the latter setting, the opposite holds for French (fr) reports. In contrast, RTLP consistently generates more diverse reports ($\downarrow$ Self-BLEU) in the multilingual setting than in the monolingual setting. For example, when generating English (en) reports, RTLP achieves Self-BLEU ≈ 38 and ≈ 52, respectively. Such a finding also contributes to the *blessing of multilinguality*.

7 Discussion

In this paper, we proposed a neural multilingual cardiac signal captioning framework. Motivated by the empirical benefits of multilingual representation learning, we designed a discriminative multilingual pre-training method, entitled RTLP. This involved randomly replacing tokens in a caption with those from a different language and tasking the network with identifying the language of all tokens. When evaluated in the multilingual setting, RTLP performed on par with state-of-the-art pre-training methods such as MLM and MARGE. We also showed that RTLP-generated reports correctly capture the pathology (disease) of ECG signals and are more reliable than those generated by MLM. Lastly, we found that with RTLP, multilingual fine-tuning is preferable to its monolingual counterpart, an observation we termed the blessing of multilinguality. Based on these findings, we recommend exploiting RTLP for multilingual cardiac signal captioning.

Looking forward, several exciting paths remain. For example, designing a captioning framework that incorporates multiple modalities (e.g., medical videos, electronic health record data) from different medical domains may mimic a more realistic clinical setting. As such, the framework may generate reports of a higher quality and increase its likelihood of adoption by medical professionals. Furthermore, generating reports of a higher level of complexity would be a suitable next step to validate the applicability of our framework.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), pp. 7674–7684, 2020.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Siddharth Biswal, Cao Xiao, M Brandon Westover, and Jimeng Sun. Eegtotext: learning to write medical reports from eeg recordings. In *Machine Learning for Healthcare Conference*, pp. 513–531. PMLR, 2019.

Siddharth Biswal, Cao Xiao, Lucas M Glass, Brandon Westover, and Jimeng Sun. Clara: Clinical report auto-completion. In *Proceedings of The Web Conference 2020*, pp. 541–550, 2020.

David J Brailer, Eugene Kroch, and Mark V Pauly. The impact of computer-assisted test interpretation on physician decision making: the case of electrocardiograms. *Medical Decision Making*, 17(1):80–86, 1997.

Richard A Caruana. Multitask connectionist learning. In *In Proceedings of the 1993 Connectionist Models Summer School*. Citeseer, 1993.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pp. 7059–7069, 2019.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, 2018.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020a.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, 2020b.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Alex Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012.

Sadid A Hasan, Yuan Ling, Joey Liu, Rithesh Sreenivasan, Shreya Anand, Tilak Raj Arora, Vivek Datla, Kathy Lee, Ashequl Qadir, Christine Swisher, et al. Attention-based medical caption generation with image modality classification and clinical concept mapping. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 224–230. Springer, 2018.

Judith H Hibbard, Ellen Peters, Paul Slovic, Melissa L Finucane, and Martin Tusler. Making health care quality reports easier to use. *The Joint Commission journal on quality improvement*, 27(11): 591–604, 2001.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. *arXiv preprint arXiv:1909.00964*, 2019a.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019b.

Qiao Jin, Bhuwan Dhingra, William W Cohen, and Xinghua Lu. Probing biomedical embeddings from language models. *NAACL HLT 2019*, pp. 82, 2019.

Alla Keselman and Catherine Arnott Smith. A classification of errors in lay comprehension of medical documents. *Journal of biomedical informatics*, 45(6):1151–1163, 2012.

Pavel Kisilev, Eli Sason, Ella Barkan, and Sharbell Hashoul. Medical image captioning: Learning to describe medical image findings using multi-task-loss cnn. *Deep Learning for Precision Medicine, Riva del Garda, Italy*, 2016.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. Pre-training via paraphrasing. *Advances in Neural Information Processing Systems*, 33, 2020.

Jiwei Li and Dan Jurafsky. Mutual information and diverse decoding improve neural machine translation. *arXiv preprint arXiv:1601.00372*, 2016.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. *arXiv preprint arXiv:1904.02633*, 2019a.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019b.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pp. 13–23, 2019.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert. Massively multilingual asr: 50 languages, 1 model, 1 billion parameters. *Proc. Interspeech 2020*, pp. 4751–4755, 2020.

David Richley and Harriet Walters. Clinical guidelines by consensus recommendations for ECG reporting standards and guidance., 2020. URL `https://scst.org.uk/wp-content/uploads/2020/06/CS4v1.1-ECG_Reporting_Guidelines_June-2020.pdf`.

Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1519–1528, 2020.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7464–7473, 2019.

C Peter Waegemann, Claudia Tessier, Andrew Barbash, Barry H Blumenfeld, Jonathan Borden, Robert M Brinson Jr, Todd Cooper, Peter L Elkin, J Michael Fitzmaurice, Susan Helbig, et al. Healthcare documentation: A report on information capture and report generation. *Newton, MA: Medical Records Institute*, 2002.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset, 2020. URL `https://physionet.org/content/ptb-xl/1.0.1/`.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9049–9058, 2018.

Wonjin Yoon, Jinhyuk Lee, Donghyeon Kim, Minbyul Jeong, and Jaewoo Kang. Pre-trained language model for biomedical question answering. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 727–740. Springer, 2019.

Xianhua Zeng, Li Wen, Banggui Liu, and Xiaojun Qi. Deep learning for ultrasound image caption generation based on object detection. *Neurocomputing*, 392:132–141, 2020.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1628–1639, 2020a.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020b.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 1097–1100, 2018.

A DATASETS

A.1 DATA PREPROCESSING

The ECG frames consisted of 2500 samples and consecutive frames had no overlap with one another. Data splits were always performed at the patient-level.

PTB-XL (Wagner et al., 2020). Each ECG recording was originally 10 seconds with a sampling rate of 500Hz. We extract 5-second non-overlapping segments of each recording generating frames of length 2500 samples. We follow the diagnostic class labelling setup suggested by Strodthoff et al. (2020) which resulted in five classes: Conduction Disturbance (CD), Hypertrophy (HYP), Myocardial Infarction (MI), Normal (NORM), and Ischemic ST-T Changes (STTC). Furthermore, we only consider ECG segments with one label assigned to them. The ECG frames were standardized to follow a standard Gaussian distribution.

A.2 DATA SAMPLES

In this section, we outline the number of instances used during training.

Table 2: Number of instances (number of patients) used during training. These represent sample sizes for all 12 leads.

Dataset	Train	Validation	Test
PTB-XL	22,670 (11,335)	3,284 (1,642)	3,304 (1,152)

A.3 VOCABULARY TOKENS

In this section, we outline the number of language-specific tokens available in each language's vocabulary for the two datasets.

Table 3: Number of language-specific tokens in each dataset

Dataset	de	el	en	es	fr	it	pt
PTB-XL	2206	2662	1606	1950	1974	1866	2010

B TRANSLATION DETAILS

In this section, we outline the steps taken to translate the ECG reports originally found in the PTB-XL dataset. We remind readers that although these ECG reports are a mixture of English and German, they are predominantly in the latter. As a result, we treat German as the source language from which we translate the reports to other languages. More specifically, we follow these steps.

1. We leverage the Google Translate API[2] to first detect the source language of each ECG report. Although the majority of the reports are in German, some are in English, and this language detection step ensures that the ultimate translation is of a higher quality.
2. We continue to leverage the Google Translate API to translate ECG reports from the identified source language to the target language of interest.
3. Due to imperfections in the Google Translate API, certain ECG reports may not be translated in full or translated at all. To minimize the incidence of such cases, we repeat Step 2 several times and stop once we reach the following criterion: we deploy the language detection module of the Google Translate API on the translated reports to confirm that over 90% of them are indeed in the translated language. Although this implies that the final translated reports may have some noise, we found that this did not prevent our algorithm from learning appropriately.

[2] https://pypi.org/project/googletrans/

C IMPLEMENTATION DETAILS

C.1 NETWORK ARCHITECTURES

In this section, we outline the neural network architectures used for our encoder and decoder. More specifically, we use the architecture shown in Table 4 for the encoder and that shown in Table 5 for the decoder.

Table 4: Encoder architecture used for experiments conducted on the PTB-XL dataset. P, C_{in}, and C_{out} represent the kernel size, number of input channels, and number of output channels, respectively. A stride of 3 was used for all convolutional layers. M represents the dimension of the final representation. We only use layer 5 when performing supervised pre-training. When captioning, layer 4 outputs T temporal features.

Layer Number	Layer Components	Kernel Dimension
1	Conv 1D BatchNorm ReLU MaxPool(2) Dropout(0.1)	$7 \times 12 \times 32$ ($P \times C_{\text{in}} \times C_{\text{out}}$)
2	Conv 1D BatchNorm ReLU MaxPool(2) Dropout(0.1)	$7 \times 32 \times 64$
3	Conv 1D BatchNorm ReLU MaxPool(2) Dropout(0.1)	$7 \times 64 \times 128$
4	Linear ReLU	$128 \times M$
5	Linear	$M \times$ C (classes)

Table 5: Decoder architecture used for experiments conducted on the PTB-XL dataset. $E = 300$ represents the dimension of the representations from the encoder and the representations of the decoder tokens. $H = 4$ represents the number of heads used in each of the self and cross-attention modules. C_{lang} represents the number of tokens in a specific language.

Layer Number	Layer Components	Kernel Dimension
1	Transformer Decoder Layer	E, H
2	Transformer Decoder Layer	E, H
3	Transformer Decoder Layer	E, H
4	Transformer Decoder Layer	E, H
5	Linear	E x C_{lang}

Table 6: Batchsize and learning rates used for training. The Adam optimizer was used for all experiments.

Stage	Batchsize	Learning Rate
Encoder		
Supervised Pre-training	128	10^{-5}
Decoder		
MLM Pre-training	128	10^{-3}
ELECTRA Pre-training	128	10^{-3}
RTLP Pre-training	128	10^{-3}
MARGE Pre-training	64	10^{-4}
Combined		
Fine-tuning	128	10^{-3}

C.2 Encoder Pre-training

In this section, we outline the task used to pre-train the encoder of the captioning system in a supervised manner. Specifically, we learn an encoder, $f_\theta : u \in \mathbb{R}^{P \times D} \rightarrow y \in \mathbb{R}^C$ parameterized by θ, that maps $P = 12$ D-dimensional ECG signals, u, (where P represents the number of leads) to a C-dimensional output representing the probability assigned to each of the cardiac arrhythmia classes. When leveraging the PTB-XL dataset, $C = 5$. For a mini-batch of size, B, and where c_i represents the ground-truth class for a particular instance, x_i, we learn this behaviour by optimizing the following categorical cross-entropy loss.

$$\mathcal{L}_{\mathcal{CE}} = -\frac{1}{B} \sum_{i=1}^{B} \log p_\theta(y_i = c_i) \tag{4}$$

We checkpoint, and eventually exploit, the parameters, θ, that coincide with the lowest loss observed on the validation set. This ensures that we use parameters that do not exhibit overfitting.

C.3 BASELINE IMPLEMENTATIONS

Masked Language Modelling. Masked language modelling (MLM) can be thought of as analogous to a denoising autoencoder. Inputs are perturbed and the network is tasked with generating the original, unperturbed version of the input. In the context of natural language processing, a fraction $F = 0.15$ of the tokens in a sentence are chosen to be masked. Of these chosen tokens, 80% are replaced with the token $[\text{MASK}]$, 10% are replaced with a random token from the vocabulary, and the final 10% are not replaced at all. The motivation behind this task lies in the ability of the network to leverage the context of masked tokens to correctly predict them. This, in turn, allows for the learning of rich representations. In our context, and to allow for a fair comparison to the multilingual pre-training methods, we follow the original implementation introduced by Devlin et al. (2019) for each of the language mini-batches. More specifically, at each iteration, we load N mini-batches corresponding to N languages and perform MLM on each of these batches.

ELECTRA. ELECTRA, as opposed to MLM introduced above, is a discriminative language representation learning method. ELECTRA builds upon the implementation of MLM in the following ways. First, instead of masking tokens and tasking the network with generating the original token, ELECTRA performs a binary classification of whether a token was replaced or not. The motivation for doing so lies in the alleged unnecessary complexity associated with generative language representation learning methods. Moreover, instead of replacing tokens with the $[\text{MASK}]$ token, ELECTRA proposes to do so by exploiting the predicted outputs of an MLM. This increases the likelihood that replaced tokens are in-distribution. As a result, ELECTRA simultaneously trains an MLM network and a binary classifier. In our context, we follow the original implementation introduced by Clark et al. (2020) for each of the N mini-batches.

MARGE. MARGE is a generative multilingual language representation learning method that exploits source documents in various languages to generate text from a similar yet distinct target document. For example, M source documents with $M * S$ tokens are encoded and leveraged by a decoder to generate the T tokens in the target document. In doing so, the network is able to capture relationships between languages and thus learn representations useful for downstream multilingual tasks. In the original implementation (Lewis et al., 2020), similar documents need to be retrieved from a database. In our context, however, our ECG reports are available in N different languages and thus the target document is formed by a report in one language and the source documents are formed by reports in the remaining $N - 1$ languages. Since our ECG reports were translated from a single original language, we used reports in this language as target documents. For PTB-XL, this amounts to using German.

C.4 EVALUATION METRICS

As we are mainly interested in the cardiac captioning task, we leverage three automatic metrics commonly used to evaluate image-captioning (BLEU score (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005) and $\text{ROUGE}-\text{L}$ (Lin, 2004)). At a high-level, these metrics quantify the degree of overlap of n-grams between a ground-truth sentence and a generated sentence. An n-gram can be thought of as a combination of tokens (words) that neighbour one another. For example, a 1-gram simply consists of all individual tokens in a sentence whereas a 2-gram consists of all pairs of adjacent tokens in the sentence.

BLEU Score. The BLEU score first requires calculating the precision of n-grams for a particular value of n. This precision is defined as the number of overlapping n-grams between the generated and ground-truth sentence, $Overlap_n$, divided by the total number of n-grams in the generated sentence $Total_n$. Such a calculation is repeated for multiple values of $n \in [1, \ldots, N]$ before being averaged and weighted according to a brevity penalty, BP, which penalizes generated sentences which are shorter than the ground-truth sentence.

$$\text{BLEU} - \text{N} = BP \cdot \left(\Pi_{n=1}^{N} Precision_n\right)^{\frac{1}{N}} \qquad Precision_n = \frac{Overlap_n}{Total_n} \tag{5}$$

METEOR Score. The METEOR score was designed, for the most part, to explicitly account for recall in n-gram overlap calculations, an operation not included in the BLEU score. Specifically, it aligns the ground-truth and generated sentences with one another, calculates the *unigram* (1-

gram) precision and recall, and derives an F-score with greater emphasis on recall than on precision. Similar to BLEU, it weights the F-score with a sentence brevity penalty, BP.

$$\text{METEOR} = \frac{10 \cdot Precision \cdot Recall}{9 \cdot Precision + Recall} \tag{6}$$

ROUGE Score. Although there exist multiple ROUGE scores (e.g., ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S), we opt for the ROUGE-L score because it obviates the need to set a value of n for the n-grams. Formally, ROUGE-L calculates the F-score (geometric mean of precision and recall) of the longest common sub-sequence (LCS) of tokens between the ground-truth sentence with m tokens and the generated sentence with n tokens.

$$\text{ROUGE} - \text{L} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{7}$$

$$Precision = \frac{LCS}{n} \quad \text{and} \quad Recall = \frac{LCS}{m}$$