

СОДЕРЖАНИЕ

Введение	3
1 Проблема интерпритации нейронных сетей	4
1.1 Методы интерпретации глубоких нейронных сетей	7
1.1.1 Local Interpretable Model-Agnostic Explanations	7
1.1.2 Deep Visualization Framework	9
1.1.3 Улучшение деконволюции: Guided backpropogation	13
1.1.4 Grad Cam	15
2 Проблема классификации ЭКГ	18
2.1 Обзор проблемы	18
2.2 Обзор датасета	19
2.3 Методы решения	20
2.4 Визуализация работы классификатора ЭКГ сигналов	22
2.5 Архитектурные эксперименты: ConvAttantionNet	29
2.5.1 Описание архитектуры	29
2.5.2 Результаты экспериментов	30
3 Вывод	38
Список использованных источников	39

Интерпретация узлов нейронной сети как семантических понятий

ВВЕДЕНИЕ

Нейронные сети достигают хороших результатов в большом количестве прикладных задач, таких как классификация изображений, обработка и анализ текста и др., широко распространены и значительно превосходят детерминированные алгоритмы в решении широкого набора задач. Однако ценой высокого качества работы нейронных сетей является низкая интерпретируемость результатов их работы. Задача интерпретации работы НС и, в частности, их отдельных узлов - актуальная проблема области исследования искусственного интеллекта. В качестве целевой задачи, решаемой НС, рассматривается задача классификации сигналов электрокардиографа на предмет наличия заболеваний. Рассматриваются методы решения этой задачи и предлагается способ интерпретации работы сети методами, используемыми в области компьютерного зрения. Кроме того, предлагается новая архитектура для решения целевой задачи, рассчитанная на повышенную интерпретируемость результатов классификации. Код доступен по ссылке <https://github.com/dupeljan/ecg-with-explanations>

1 Проблема интерпритации нейронных сетей

Интерпретация - это отображение абстрактного понятия в понятную для человека область Объяснение - это набор характеристик интерпретируемой области, которые привели к принятию данного решения в конкретном случае [1]

Степень доверия к результатам работы классификатора является основной характеристикой во время принятия того или иного метода для решения реальных прикладных задач. Модели, не имеющие большого доверия со стороны пользователей, редко используются на практике. Можно различать два разных (но связанных) определения доверия: (1) доверие к прогнозу, т. е. доверяет ли пользовательциальному прогнозу в достаточной степени, чтобы предпринять какое-либо действие на его основе, и уровень доверия к модели: факт того, что модель будет вести себя разумно во время работы над реальными прикладными задачами. Определение доверия к индивидуальным прогнозам - важная проблема, когда модель используется для принятия решений. Например, при использовании машинного обучения в медицинской диагностике или детекции актов терроризма, прогнозы не могут быть основаны на слепой вере в модель, поскольку последствия могут быть катастрофическими.

Помимо доверия к индивидуальным прогнозам, также полезно оценивать модель в целом перед ее развертыванием. Чтобы допустить ту или иную НС, пользователи должны быть уверены в том, что модель будет хорошо работать на реальных данных, согласно интересующей их метрике. В настоящее время модели оцениваются с использованием метрик точности в доступном заранее наборе тестовых данных. Однако реальные данные часто значительно отличаются от тестовых, и, кроме того,

целевая метрика может не учитывать реальные цели прикладной задачи.

Объяснение причин помогает увеличить степень доверия пользователей к предсказаниям модели. Зачастую тренировочные и валидационные данные включают в себя некоторые признаки, не имеющиеся в реальных данных. Например, в случае классификации МРТ снимков тренировочный набор может состоять из уже размеченных врачами снимков, с пометками, которые очень сильно коррелируют с диагнозом. Метод объяснения LIME способен отследить такое переобучение. Кроме того, объяснения предсказаний набора моделей может сыграть важную роль в решающем выборе архитектуры для решения задачи.

Важный критерий объяснения работы классификатора - их интерпретируемость т.е качественное объяснение взаимосвязи входа и выходе модели.

В результате развития области искусственного интеллекта модели классификации усложняются до такой степени, что объяснение причины, по которой алгоритм принял то или иное решение является сложной задачей. Хотя самые первые системы искусственного интеллекта были легко интерпретируемыми, в последние. Эмпирический успех моделей глубокого обучения является результатом комбинации эффективных алгоритмов обучения и их огромного параметрического пространства. Новейшие нейронные сети имеют сотни слоев и миллионы параметров, что позволяет рассматривать глубокие нейронные сети как черный ящик - его работа абсолютно нетривиальна. Противоположностью «черного ящика» является прозрачность, то есть поиск прямого понимания механизма работы модели.

По мере того как модели машинного обучения все чаще

используются для важных прогнозов в критически важных для человечества областях, требования к прозрачности со стороны различных заинтересованных сторон в ИИ возрастают. Главная опасность заключается в создании и использовании решений, которые не являются оправданными, законными или просто не позволяют получить подробные объяснения их поведения. Пояснения, поддерживающие выходные данные модели, имеют решающее значение, например, в точной медицине, где экспертам требуется гораздо больше информации от модели, чем простое двоичное предсказание для подтверждения своего диагноза. Среди других примеров - автономные транспортные средства в сфере транспорта, безопасности и финансов.

Авторы статьи [2] выделяют три основных мотивации в пользу развития алгоритмов интерпретации нейронных сетей:

- Интерпретируемость помогает обеспечить беспристрастность при принятии решений, то есть обнаруживать и, следовательно, исправлять предвзятость в наборе обучающих данных.
- Интерпретируемость способствует обеспечению надежности, выделяя потенциальные нежелательные возмущения, которые могут изменить прогноз.
- Интерпретируемость может действовать как гарантия того, что только значимые переменные определяют результат, то есть гарантируют, что в рассуждениях модели существует истинная причинно-следственная связь.

1.1 Методы интерпретации глубоких нейронных сетей

1.1.1 Local Interpretable Model-Agnostic Explanations

Метод Local Interpretable Model-Agnostic Explanations (LIME)[2] позволяет получить точечную интерпретацию любого классификатора независимо от его архитектуры. Для каждого отдельного класса классификационная модель определяется как функция $f : R^n \rightarrow R$, определенная на множестве признаков и принимающая значение уверенности в том, что её аргумент является элементом выбранного класса. Элемент выборки, результат классификации которого необходимо объяснить, определяется как $x, x \in R^d$. Интерпретация примера x определяется как $x', x' \in \{0, 1\}^{d'}$ - некоторым образом введенная бинарная маска x . Как видно, d и d' могут не совпадать. Так, например, для изображений можно использовать попиксельную бинаризацию $d' = \frac{d}{3}$. Вводится понятие интерпретируемой модели $g \in G$, где G - класс классификационных легко интерпретируемых моделей, таких как линейная модель, дерева решений и др. $\forall g \in G g : \{0, 1\}^{d'} \rightarrow R$ где значение функции g уверенность модели в правильной классификации. Для того, чтобы регулировать простоту интерпретации модели используется функция $\Omega(g)$, $g \in G$ - сложность функции g . Для каждого класса функция Ω может определяться по-разному: для линейной функции это может быть количество параметров, а для дерева решений - количество ветвей. Для работы алгоритма также требуется функция расстояния между элементами выборки $\pi_x(z)$. Для выбора оптимальной модели оптимизируется функция $\mathcal{L}(f, g, \pi_x)$, являющаяся мерой близости интерпретируемой функции g к реальному классификатору f по мере близости их аргумента π_x . Для того, чтобы приблизить функции f и g ,

генерируются все возможные элементы $Z' = z' : z' \in \{0, 1\}^{d'}$, после чего получается множество Z как результат маскирования исходного элемента x масками из Z' . Обозначим за $\mathcal{L}_Z(f, g, \pi_x)$ меру разности функций $\mathcal{L}(f, g, \pi_x)$ на множестве Z . Тогда интерпретация входа x есть функция $\xi(x)$:

$$\xi(x) = \arg \min_{g \in G} \mathcal{L}_Z(f, g, \pi_x) + \Omega(g)$$

В своей статье [2] для конкретной реализации были выбраны следующие элементы:

$$G = g(z') : g(z') = w_g \cdot z', w_g \in R^{d'}$$

$$\pi_x(z) = \exp(-D(x, z)^2 / \sigma^2)$$

$$\mathcal{L}_Z(f, g, \pi_x) = \sum_{z, z' \in Z} \pi_x(z)(f(z) - g(z'))^2$$

где D - мера расстояния между x и z , для изображений L_2 норма. Функция Ω , штраф за сложность, определяется как $\Omega = \infty \cdot [\|w_g\|_0 > K]$ где K - уровень сложности интерпретации. Так как Ω не дифференцируемая функция, в качестве метода приближения w_g авторы используют модифицированный метод Lasso, оставляющий только K элементов из w_g .

Для проверки работы алгоритма LIME авторы применили этот метод для двух классификаторов: SVM классификация текстов и InceptionNet классификация изображений. В первом случае благодаря интерпретации авторы смогли обнаружить сильную корреляцию между некоторыми словами и классами, что заставляло сеть опираться только на эти ключевые слова. Пример интерпретации работы InceptionNet изображен на рисунке 1.1 .

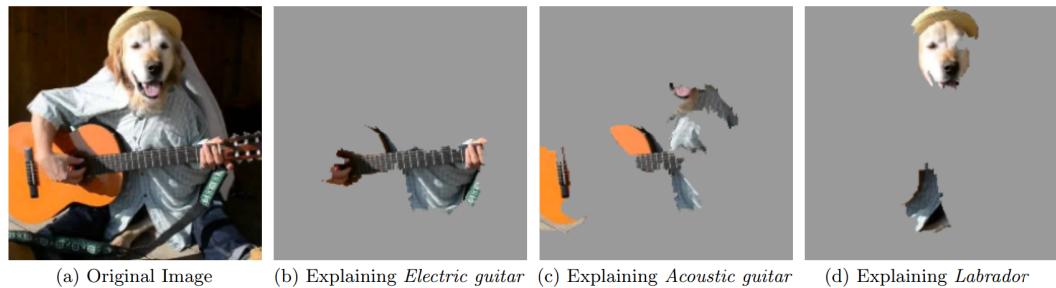


Рисунок 1.1 – Пример интерпретации работы InceptionNet при помощи метода LIME. Для интерпретации выбраны топ 3 класса, определенные моделью на изображении: ”Electric Guitar”($p = 0.32$), ”Acoustic guitar”($p = 0.24$) и ”Labrador”($p = 0.21$) где p - уверенность модели в классе от 0 до 1

Неоспоримым достоинством метода LIME является его независимость от задачи метода классификации. Общие формулы позволяют определить функции $G, \Pi(x), \mathcal{L}(f, g, \pi_x)$ так, чтобы метод давал лучший результат с точки зрения интерпретации. Однако главное преимущество этого метода также является и его недостатком, так как опуская специфику работы классификатора метод теряет большое количество информации, которая потенциально могла бы улучшить качество интерпретации. Кроме того, к недостаткам можно отнести и время работы, которое можно описать как $O(z \cdot f)$ где z - мощность множества Z' а f - сложность работы классификатора. Как заявляют авторы, для интерпретации одного изображения в их примере у них уходило около 10 минут [2].

1.1.2 Deep Visualization Framework

Jason Yosinski и др. в 2015 году представили Deep Visualization Framework[3] - программу визуализации работы сверточных нейронных сетей. Для визуализации использовались следующие техники:

- Визуализация активаций каждого конволовиционного слоя
- Деконволюция
- Градиентный подъем

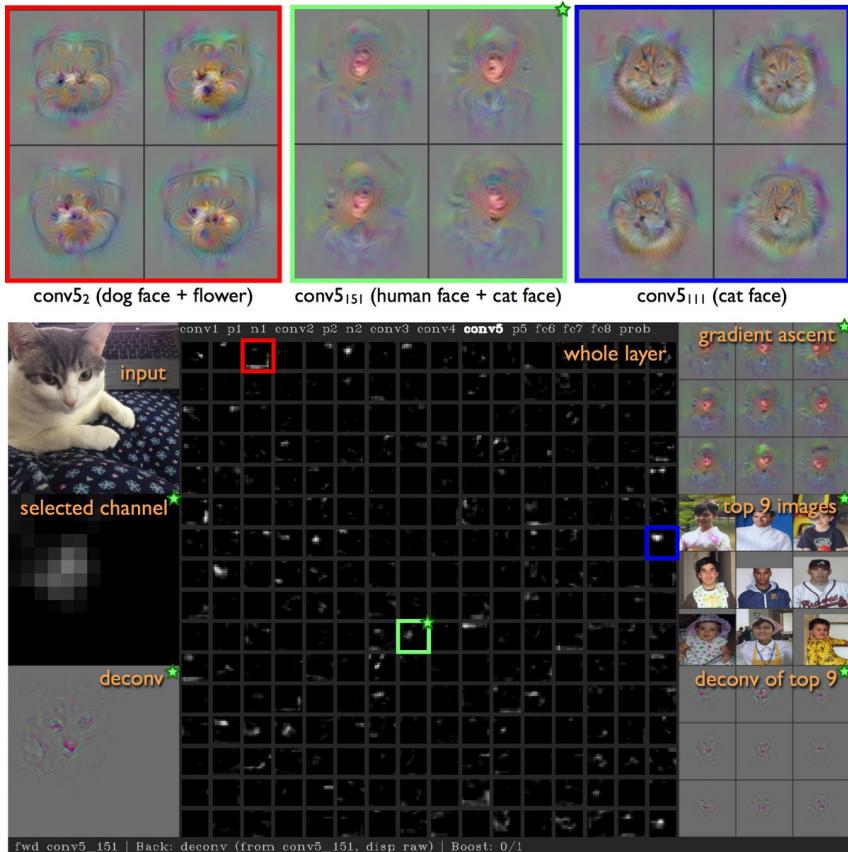


Рисунок 1.2 – Интерфейс программы Deep Visualization Framework.

Визуализация активаций

Каждый конволовиционный и полносвязный слой Н.С. в результате работы передает следующему слою Feature Map, тензор вида (B, H, W, C) для сверточ и (B, C) для полносвязных слоев, где B - размер батча, H, W - ширина и высота соответствующей feature map и C - количество каналов. Для визуализации используется одно изображение ($B = 1$) и карты активаций $(1, H, W, i)$ визуализируются для всех $i = \overline{1, C}$. На рисунке 1.2 область визуализации активаций обозначена как whole layer. Интерпретировать feature map можно как геометрическую карту, каждый

элемент которой отвечает некоторой области входного изображения. Значение feature map в какой-то точке $(1, h, w, c)$ есть результат применения канала c выбранной конволюции к области исходного изображения, отвечающей точке карты активации (h, w) . Таким образом яркие области feature map отечают областям исходного изображения, на которые среагировал заданный канал свертки. На рисунке 1.3 изображена одна из карт активаций, сопоставленная с исходным изображением. Видно, что выбранная карта активизируется на областях, на которых присутствуют лица или даже морды животных.

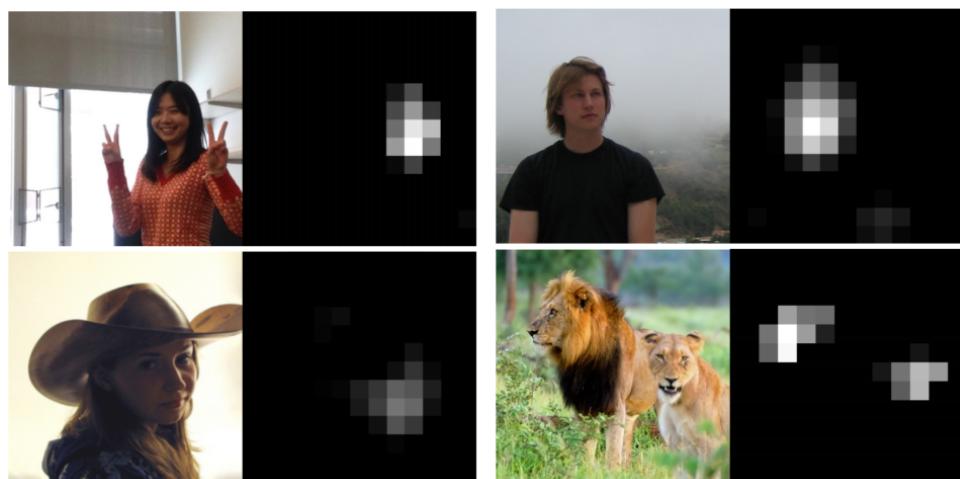


Рисунок 1.3 – Визуализация одного из каналов активаций сверточной нейронной сети, сопоставленная с исходным изображением. Видно, что выбранный канал имеет яркие области в местах, соответствующих областям лиц

Градиентный подъем

Градиентный подъем (gradient ascent) [3] - метод, строящий входное изображение, максимизирующий выход выбранного канала. Для реализации на первом шаге фиксируется входное изображение - многоканальный белый шум и фиксируются все веса исследуемой

сети. Производится тренировка сети т.е. forward pass и backward pass, но обучаемыми параметрами являются только пиксели изображения. Однако если не использовать никаких регуляризаций, изображение после тренировки сложно интерпретировать так как оно мало отличается от случайного шума. Авторы Deep Visualisation Framework предлагают набор регуляризаций, который позволяет получать более интерпретируемые изображения. В частности даже L_2 норма делает изображение гораздо понятнее для человека. На рисунке 1.2 результаты градиентного подъема обозначены как gradient ascent.

Деконволюция

Метод деконволюции[4] - один из способов реализации функции, обратной к конволюционной нейронной сети. Реализация основывается на построении деконволюционной модели, состоящей также из сверток и пулинг слоев, имеющих те же веса, что и в исследуемой конволюционной модели. Однако деконволюционная модель на вход получает выход конволюционной сети и выдает изображение. Операции активации и пулинга не являются обратимыми, поэтому для реализации обратной к конволюции функции используются значения, полученные во время прямого хода модели (switches)[4]. Для maxpool слоя, например, результатом инверсии является позиция максимального значения карты признаков предыдущего слоя.

Чтобы исследовать работы конволюционной нейросети, к каждому её слою присоединяется аналог обратной функции этого слоя, таким образом обеспечивая непрерывный путь от результатов модели к пикселям изображения. Для начала, входное изображение передается исходной сети, и для каждого слоя вычисляется функция активации. Чтобы получить

обратное изображение, на последнем слое все активации, кроме целевой, зануляются, и значение активации передается в деконволюционную сеть, которая и выдает результирующее изображение. Результат деконволюции можно интерпретировать как области интереса исследуемой сети. На рисунке 1.2 деконволюция обозначена как deconv. Слева, в сравнении с исходным изображением видно, что область интереса сети для выбранного класса - морда кота. В то же время изображение клавиатуры на фоне никак не влияет на решение модели.

1.1.3 Улучшение деконволюции: Guided backpropagation

Для визуализации конволовационных нейронных сетей в качестве основы авторы[5] использовали подход деконволюции, однако привнесли некоторые улучшения. Процесс деконволюции можно описать следующим образом: На основе карты признаков исследуемого слоя, deconvnet пытается инвертировать операции каждого слоя, транслируя выход сети в исходное изображение. Однако для необратимых операций для псевдоинвертирования используются значения feature map прямого хода модели. Это делает метод деконволюции зависимым от входного изображения, и метод guided deconvolution позволяет избавиться от этой зависимости, получая более интерпретируемый результат. Это достигается путем простого удаления pooling слоев из классификационной модели. Сравнение результатов с методом деконволюции можно увидеть на рисунке 1.4

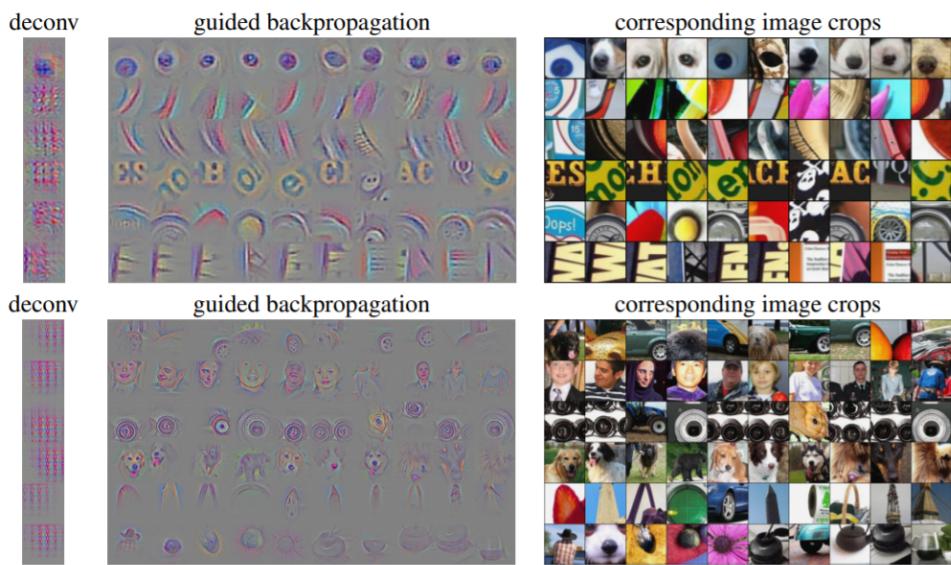


Рисунок 1.4 – Сравнение результатов визуализации guided backprop и деконволюции

Процесс визуализации guided backprop описан на рисунке 1.5. Сначала выбирается входное изображение и в сети выбирается нейрон или слой активации, работу которого необходимо визуализировать. Сеть производит прямой ход до слоя, в котором присутствует выбранная активация. В этом слое зануляются все нейроны и слои, за исключением выбранных в начале. С выбранного слоя производится градиентный спуск, однако по ходу движения назад градиенты последовательно зануляются. Обозначая за f^l - слой начала градиентного спуска, градиенты предыдущих слоев вычисляются как $\frac{\partial f^l}{\partial f^k}_{guided} = \frac{\partial f^l}{\partial f^{k+1}}_{guided} \cdot max(0, \frac{\partial f^{k+1}}{\partial f^k})$

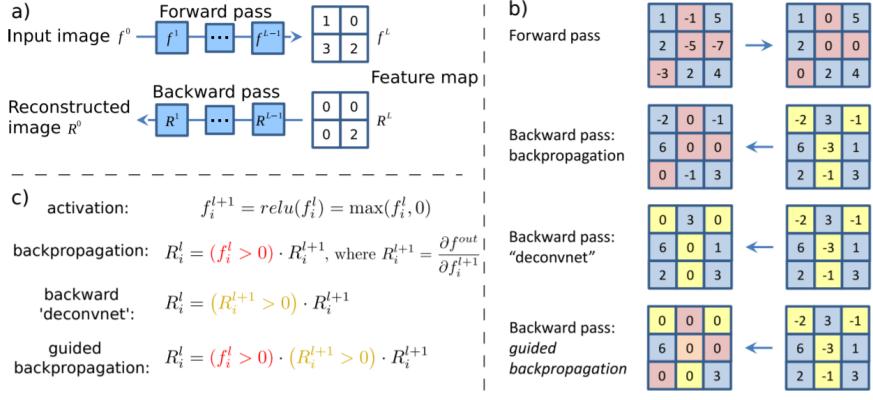


Рисунок 1.5 – Принцип работы метода guided backpropagation в отличие от деконволюции

Достоинством метода guided backprop является значительные улучшения результатов визуализации в сравнении с деконволюцией. Однако этот метод фиксирует архитектуру моделей, с которой может работать и не так эффективен как метод, описанный в следующей главе.

1.1.4 Grad Cam

Относительно современный метод визуализации работы конволовиционных сетей, не зависящий от наличия пулинг слоев и задачи классификации[6]. Gradcam предполагает наличие в конце модели полносвязного слоя, каждый нейрон которого отвечает за предсказание одного из результирующих классов. На практике большинство конволовиционных сетей, будь то сети классификации изображений или звуков, или даже сети, решающие задачу детектирования объектов, состоят из двух частей: головы (head) и шеи (backbone). Суть разделения заключается в том, что исходные данные путем прямого хода backbone части переводятся в латентное пространство признаков, являющиеся, по сути, областью определения head части модели. Такой подход удобен тем,

что для разных задач для изображений (классификация, детектирование, сегментирование) можно использовать одну и ту же backbone часть модели. Метод GradCam использует карту признаков последнего конволюционного слой backbone части модели для визуализации работы всей модели.

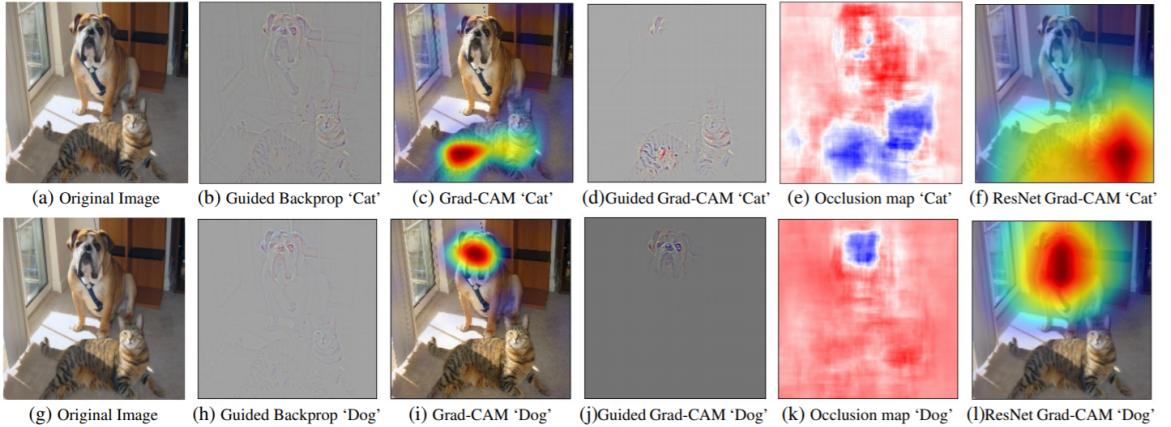


Рисунок 1.6 – Сравнение методов визуализации классификации с методом визуализации grad cam

Введем обозначения: $A^k \in R^{u \times v}, k \in K$ - карта признаков под номером k самого глубокого слоя backbone части сети, y^c - выход последнего полно связного слоя, отвечающий за класс c . Чтобы получить карту интереса сети для выбранного изображения I методом GradCam $L_{Grad=CAM}^c \in R^{u \times v}$ шириной u и высотой v для класса c необходимо произвести прямой ход модели на изображении I до финального полно связного слоя, и вычислить градиенты $\frac{\partial y^c}{\partial A_{i,j}^k}, \forall k \in K$. Далее на основе этой информации вычисляется весовой коэффициент для каждой карты признаков A^k :

$$\alpha_k^c = \frac{1}{u+v} \sum_{i=1}^u \sum_{j=1}^v \frac{\partial y^c}{\partial A_{i,j}^k}$$

Что по сути является усредненным значением градиенты в выбранной

карте признаков. Финальная визуализация вычисляется по формуле:

$$L_{Grad-CAM}^c = \text{ReLU}\left(\sum_{k \in K} \alpha_k^c A^k\right)$$

Функция ReLU используется для исключения элементов визуализации, вносящих отрицательный вес в предсказании выбранного класса. После этого полученная визуализация равномерно растягивается на область всего изображения. Кроме того, метод gradCam можно объединить с методом визуализации этого процесса можно наблюдать на изображении 1.7.

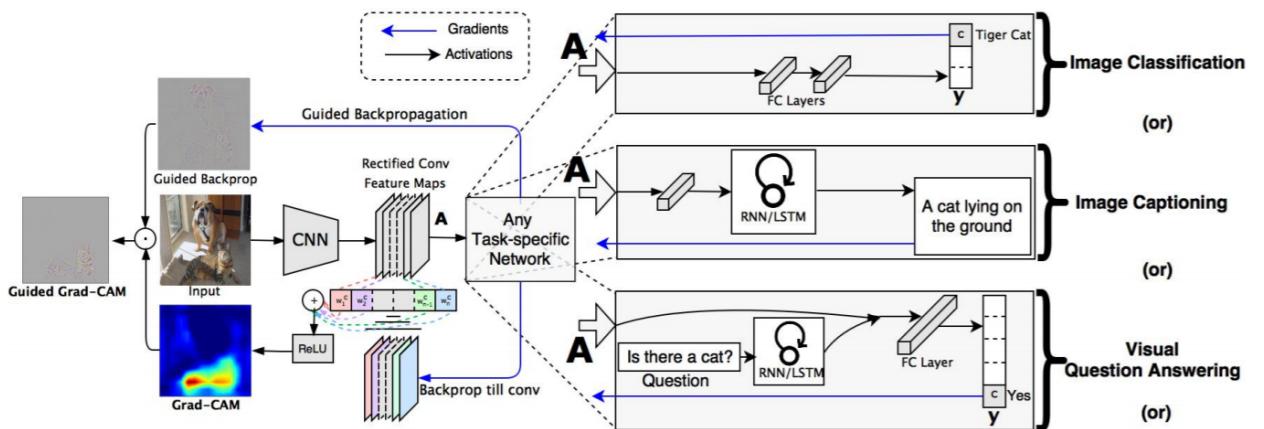


Рисунок 1.7 – Принцип работы метода GradCam

Метод визуализации GradCam прост в реализации и требует небольшого изменения архитектуры моделей, что делает его хорошим кандидатом для визуализации работы конволовиционных сетей.

2 Проблема классификации ЭКГ

2.1 Обзор проблемы

Сердечно-сосудистые заболевания являются ведущей причиной смерти во всем мире, и электрокардиограмма (ЭКГ) является важным инструментом в их диагностике. По мере перехода ЭКГ с аналоговой на цифровую, автоматизированный компьютерный анализ стандартных электрокардиограмм в 12 каналах приобрел значение в процессе медицинской диагностики. Однако ограниченная производительность классических алгоритмов исключает его использование в качестве автономного диагностического инструмента и отводит им вспомогательную роль[7].

Кратковременная стандартная ЭКГ в 12 каналов (S12L-ECG) - это наиболее часто используемое дополнительное обследование для оценки состояния сердца, применяемое во всех клинических учреждениях, от центров первичной медико-санитарной помощи до отделений интенсивной терапии. В то время как долгосрочный мониторинг сердца, например, при холтеровском обследовании, дает информацию в основном о сердечном ритме и реполяризации, S12L-ЭКГ может предоставить полную оценку электрической активности сердца. В список классифицируемых болезней входят аритмии, нарушения проводимости, острые коронарные синдромы, гипертрофия и увеличение сердечной камеры и даже эффекты лекарств и электролитные нарушения[7].

2.2 Обзор датасета

Набор данных ЭКГ PTB-XL[8] - это большой набор данных из 21837 клинических ЭКГ в 12 отведениях от 18885 пациентов длительностью 10 секунд. Необработанные данные формы волны были аннотированы двумя кардиологами, которые присвоили каждой записи потенциально несколько отчетов ЭКГ. Всего 71 отчет ЭКГ соответствует стандарту SCP-ЭКГ и охватывает диагностические, формуляры и ритмические утверждения. Чтобы обеспечить сопоставимость алгоритмов машинного обучения, обученных на наборе данных, исследователи предоставляют рекомендуемые к использованию разбиения на обучающие и тестовые наборы. В сочетании с обширными аннотациями это превращает набор данных в богатый ресурс для обучения и оценки алгоритмов автоматической интерпретации ЭКГ. Набор данных дополняется обширными метаданными по демографическим характеристикам, характеристикам инфаркта, вероятности диагностических заявлений ЭКГ, а также аннотированными свойствами сигналов.

Данные, лежащие в основе набора данных ЭКГ PTB-XL, были собраны с помощью устройств компании Schiller AG в течение почти семи лет с октября 1989 года по июнь 1996 года. С приобретением оригинальной базы данных у Schiller AG полные права на использование были переданы Physikalisch-Technische Bundesanstalt (PTB). Записи были курированы и преобразованы в структурированную базу данных в рамках долгосрочного проекта в PTB. База данных использовалась в ряде публикаций, но доступ к ней до сих пор оставался ограниченным. Комитет по институциональной этике одобрил публикацию анонимных данных в базе данных открытого доступа (PTB-2020-1). В ходе процесса публичного выпуска в 2019

году существующая база данных была оптимизирована для удобства использования и доступности сообществу машинного обучения. Данные ЭКГ и метаданные были преобразованы в открытые форматы данных, которые легко обрабатываются стандартным программным обеспечением.

2.3 Методы решения

В течение последних 22 лет PhysioNet и Computing in Cardiology совместно организовали серию ежегодных соревнований для решения клинически интересных задач, которые либо не решены, либо не решены должным образом. В последний раз соревнование проходило в 2020 году, и в качестве основной архитектуры решения задачи классификации ЭКГ использовались конволовационные сети[9].

В качестве модели классификации ЭКГ в этой работе используется модель Antônio H. Ribeiro[7], имеющаяся в открытом доступе. В качестве архитектуры была выбрана resnet подобная одномерная сеть, арихтектура которой изображена на рисунке 2.1

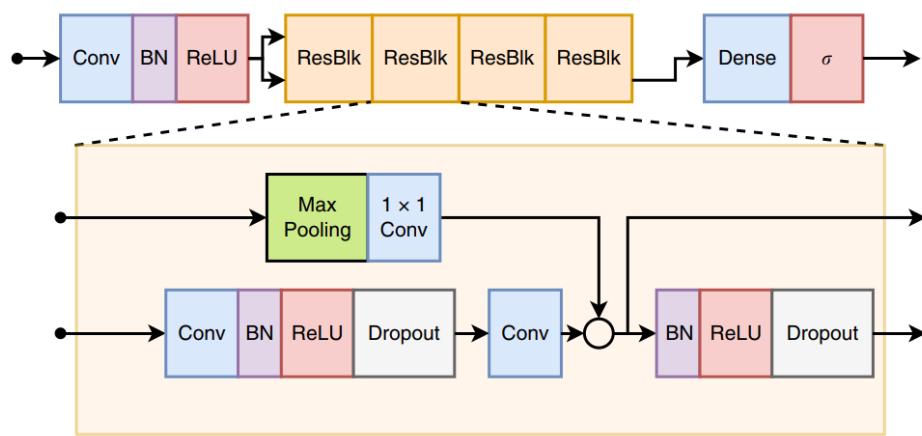


Рисунок 2.1 – Архитектура модели классификации ЭКГ, предложенная
Antônio H. Ribeiro

Table 2 (Performance indexes) Scores of our DNN are compared on the test set with the average performance of: (i) 4th year cardiology resident (cardio.); (ii) 3rd year emergency resident (emerg.); and (iii) 5th year medical students (stud.).

Precision (PPV)				Recall (Sensitivity)				Specificity				F1 score				
DNN	cardio.	emerg.	stud.	DNN	cardio.	emerg.	stud.	DNN	cardio.	emerg.	stud.	DNN	cardio.	emerg.	stud.	
1dAVb	0.867	0.905	0.639	0.605	0.929	0.679	0.821	0.929	0.995	0.997	0.984	0.979	0.897	0.776	0.719	0.732
RBBB	0.895	0.868	0.963	0.914	1.000	0.971	0.765	0.941	0.995	0.994	0.999	0.996	0.944	0.917	0.852	0.928
LBBB	1.000	1.000	0.963	0.931	1.000	0.900	0.867	0.900	1.000	1.000	0.999	0.997	1.000	0.947	0.912	0.915
SB	0.833	0.833	0.824	0.750	0.938	0.938	0.875	0.750	0.996	0.996	0.996	0.995	0.882	0.882	0.848	0.750
AF	1.000	0.769	0.800	0.571	0.769	0.769	0.615	0.923	1.000	0.996	0.998	0.989	0.870	0.769	0.696	0.706
ST	0.947	0.968	0.946	0.912	0.973	0.811	0.946	0.838	0.997	0.999	0.997	0.996	0.960	0.882	0.946	0.873

PPV positive predictive value. The bold values represent the best scores.

Рисунок 2.2 – Результаты валидации модели на тестовом датасете, предоставленном в статье

Модель имеет хорошую точность на предоставленном тестовом датасете, однако имеет ряд недостатков: модель натренирована на закрытом датасете с 6 не пересекающимися классами и использует одномерные сигналы разрешением 400Hz, хотя датасет PTB-XL предоставляет данные в разрешении 500Hz и имеет порядка 30 классов. В рамках работы модель была протестирована на 5 разбиении PTB-XL на примерах, имеющих только допустимые наборы классов для модели из этой главы. Результаты работы, представленные в таблице 2.1, показывают худший результат работы модели на новых данных.

Таблица 2.1 – Результаты валидации модели на 5 разбиении PTB-XL на примерах, имеющих только допустимые наборы классов

	precision	recall	f1-score	support
1dAVb	0.58	0.76	0.66	79
RBBB	0.85	0.93	0.88	54
LBBB	0.84	0.91	0.87	53
SB	0.89	0.66	0.76	64
AF	0.95	0.91	0.93	152
ST	0.86	0.90	0.88	83

2.4 Визуализация работы классификатора ЭКГ сигналов

В качестве метода визуализации работы выбранной модели классификации выбран метод GradCam. В качестве входных изображений были взяты ЭКГ сигналы из датасета PTB-XL со следующими порядковыми номерами: 282, 424, 489, 1694, 19715, 21585. Эти кардиограммы относятся к разным допустимым для модели классам и, кроме того, правильно классифицируются моделью. Метод GradCam предполагал наличие двумерной карты признаков, однако формула легко обобщается на одномерный случай. На рисунках 2.3- 2.8. Метод GradCam не разделяет входные каналы, поэтому карта интереса модели общая для всех 12 каналов.

Diagnose: AFIB Predicted: AFIB



Рисунок 2.3 – Области внимания модели для ЭКГ сигнала с диагнозом

”Мерцательная аритмия”

Diagnose: CRBBB Predicted: CRBBB

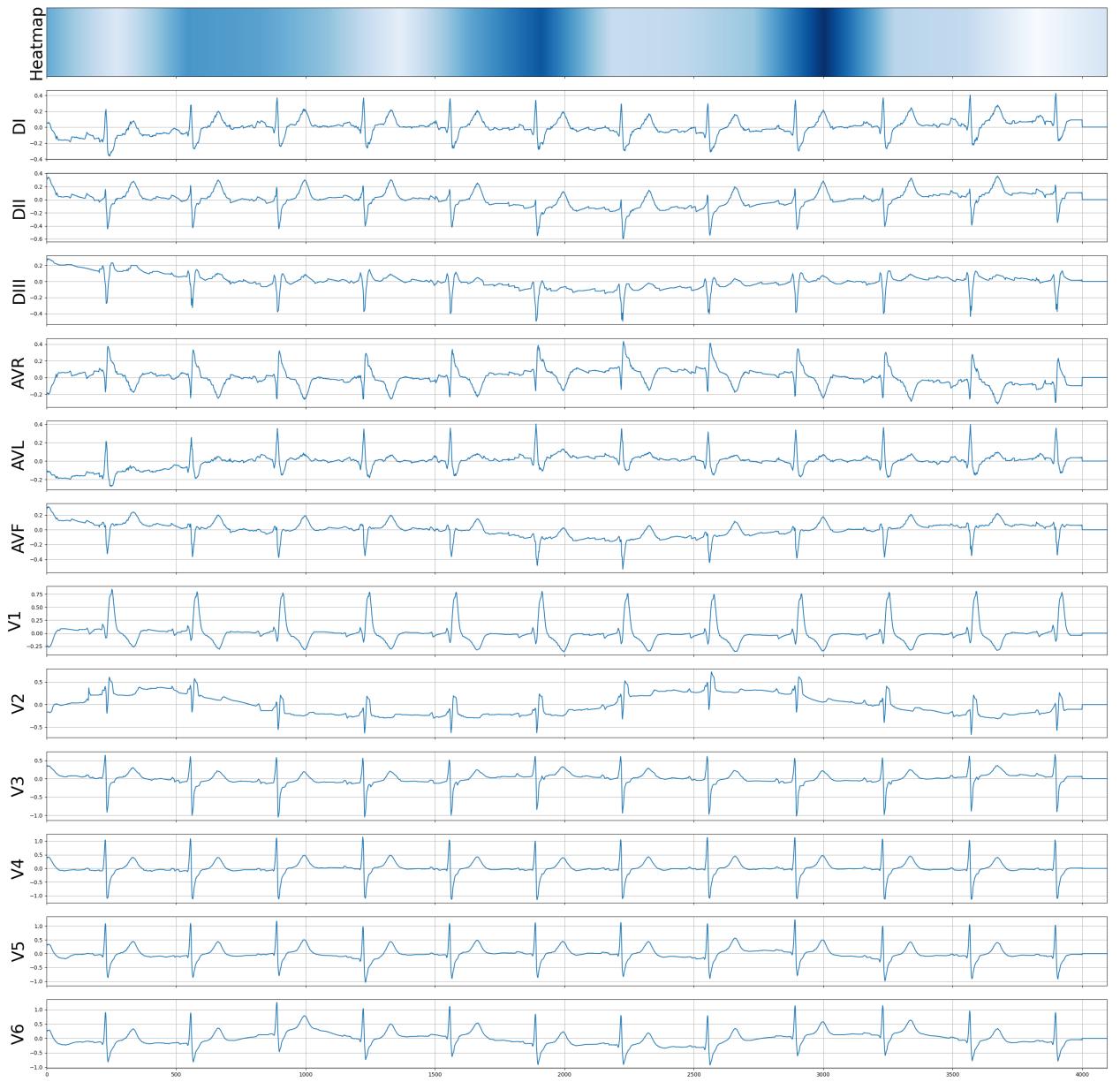


Рисунок 2.4 – Области внимания модели для ЭКГ сигнала с диагнозом

”Блокада правой ножки пучка Гиса”

Diagnose: CLBBB Predicted: CLBBB

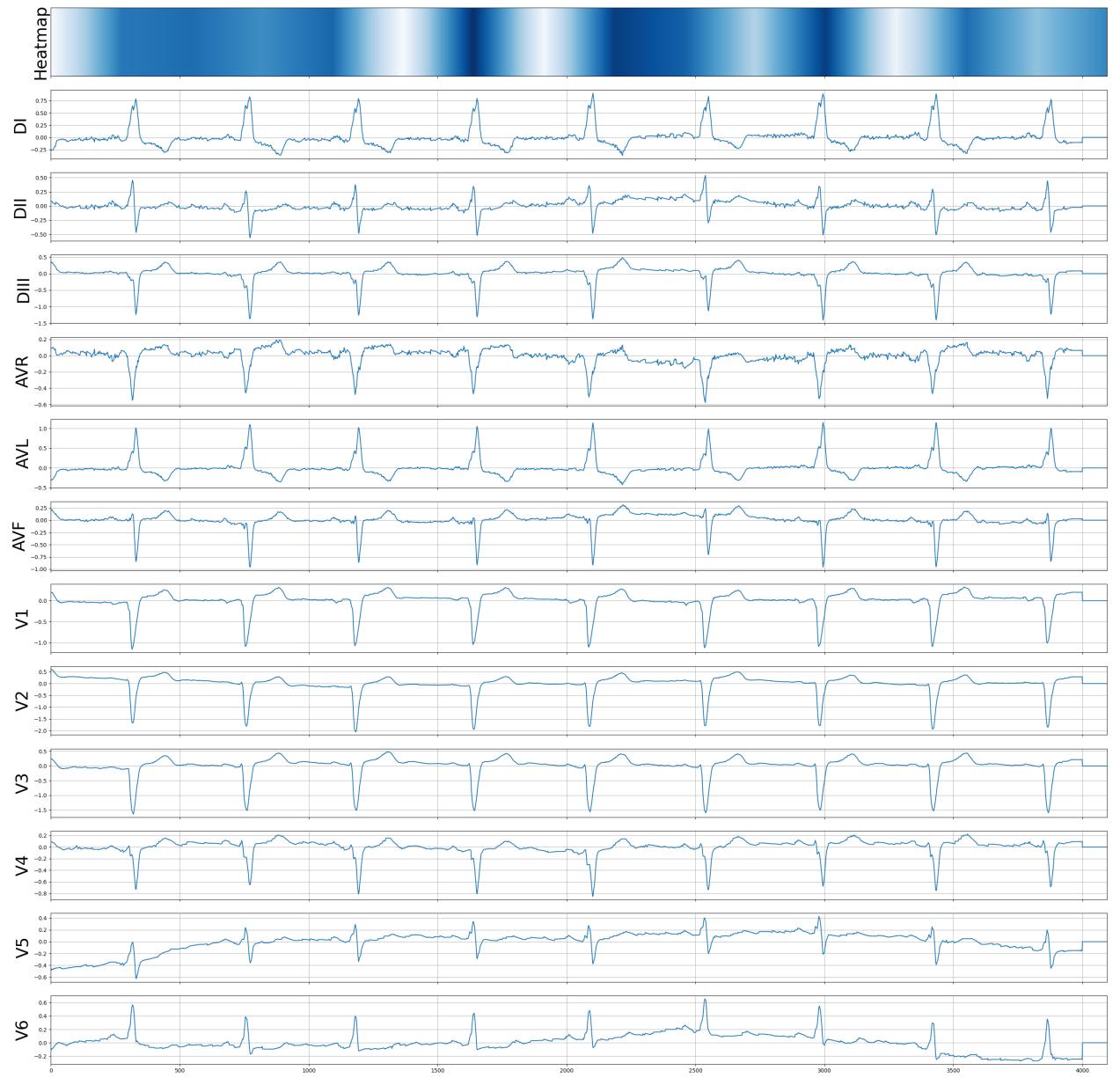


Рисунок 2.5 – Области внимания модели для ЭКГ сигнала с диагнозом
”Блокада левой ножки пучка Гиса”

Diagnose: 1AVB,CLBBB Predicted: CLBBB

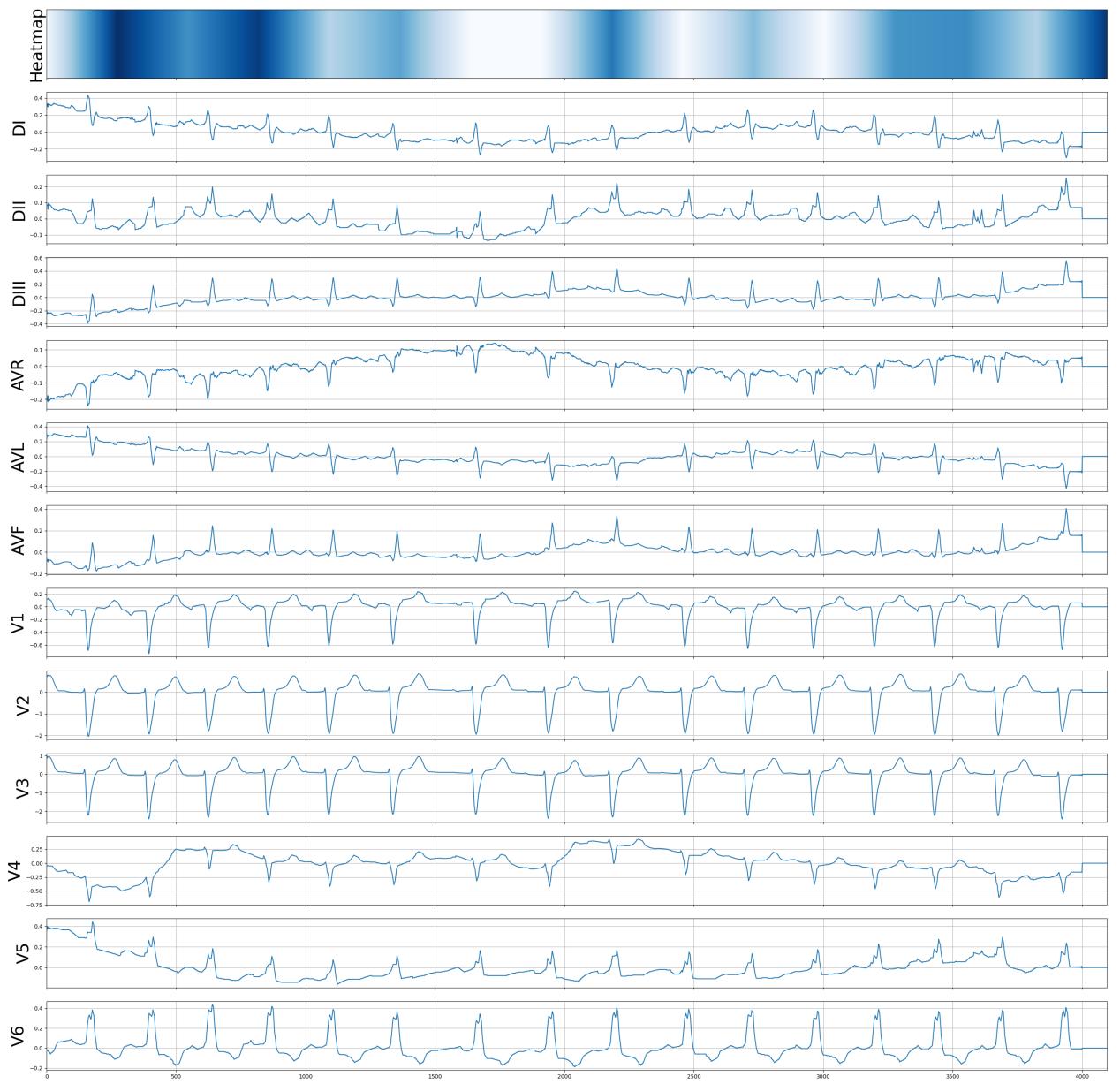


Рисунок 2.6 – Области внимания модели для ЭКГ сигнала с диагнозом
”Атриовентрикулярная блокада первой степени и блокада левой ножки
пучка Гиса ”

Diagnose: SBRAD Predicted: SBRAD

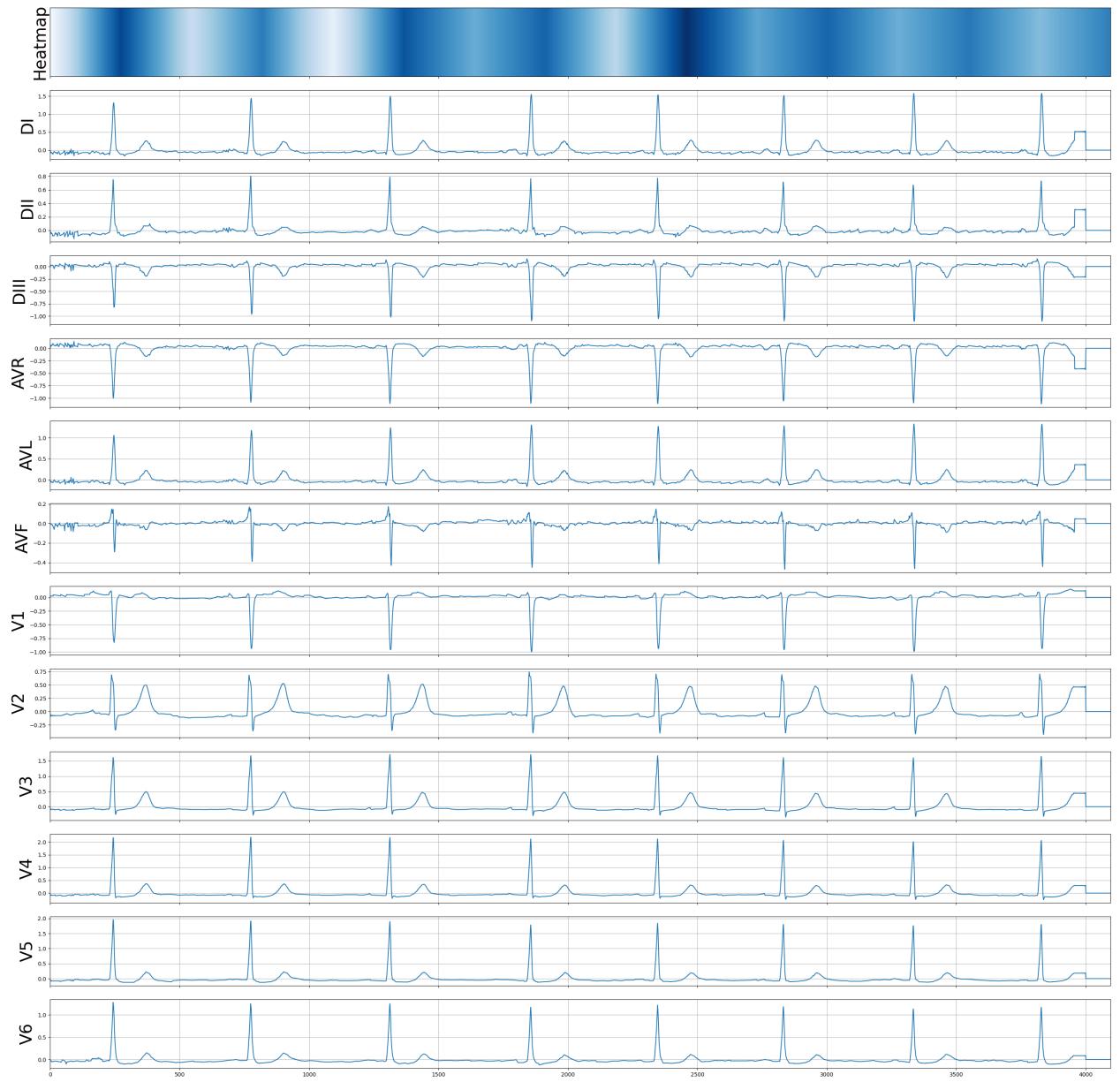


Рисунок 2.7 – Области внимания модели для ЭКГ сигнала с диагнозом
”синусовая брадикардия”

Diagnose: STACH Predicted: STACH

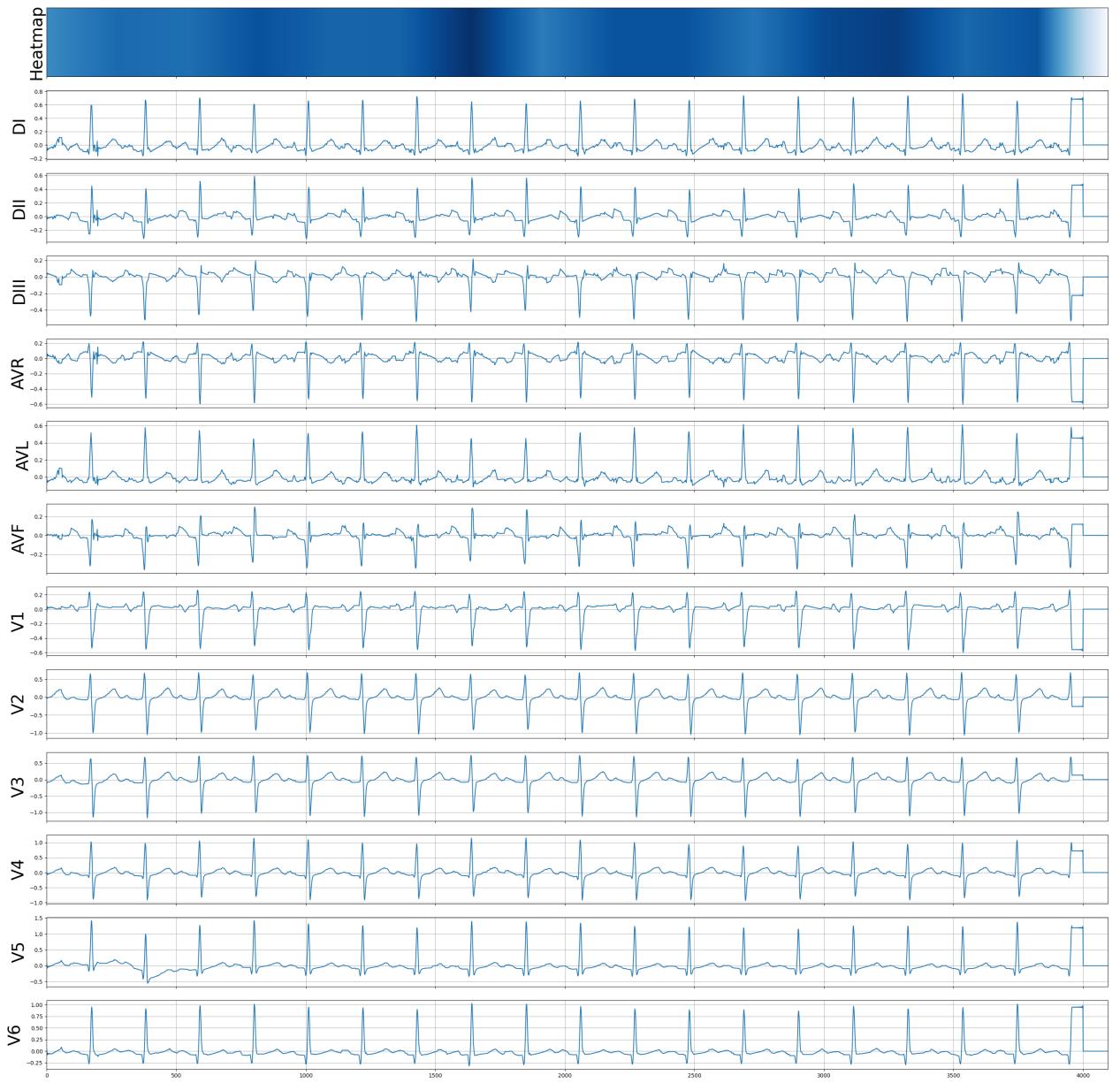


Рисунок 2.8 – Области внимания модели для ЭКГ сигнала с диагнозом

”синусовая тахикардия”

2.5 Архитектурные эксперименты: ConvAttantionNet

2.5.1 Описание архитектуры

Недостатком визуализаций конволовиционной модели resnet типа является общая карта интересов для всех каналов входного сигнала. Для того, чтобы преодолеть это ограничение, предлагается использовать новую архитектуру, реализованную в коде курсовой работы. Модель состоит из трех частей:

- Конволюционная часть, состоящая из четырех последовательных блоков 12 групп конволовий: свертка, batchNorm слой, maxPool слой;
- 6 последовательных слоев трансформер энкодеров, принимающие на вход, последовательно, 12 выходов групп конволовий из предыдущей части модели;
- Классификационная часть, состоящая из одного конволюционного блока из первой части модели, соединенного с полносвязным слоем.

Первая часть представляет из себя 12 отдельных конволовиционных сетей для каждого из каналов. В этой части сети входные каналы не имеют никакой информации друг о друге. Используя последнюю свертку этого блока как слой для применения GradCam можно получить 12 отдельных карт внимания сети для каждого из входных каналов.

Вторая часть сети выполняет роль объединения результатов свертки вместе т.е. агрегирование информации раздельно по каждому из каналов. Последний слой использует один конволюционный слой для уменьшения размерности пространства и служит в качестве финального классификатора сети.

2.5.2 Результаты экспериментов

Для тренировки модели в качестве функции потерь использовался FocalLoss[11] с параметрами $\alpha = 10, \gamma = 5$, который значительно улучшил результаты обучения в сравнении со стандартной функцией потерь MSE. В качестве алгоритма оптимизации использовался метод Adam[12], leaning rate=1e-6 с регуляризационным L_2 членом с коэффициентом 0.1. Размер батча - 256, всего тренировка требует 250 эпох. Константа обучения менялась по ходу тренировки при помощи экспоненциального планировщика(Exponential Scheduler)[12] с параметром $\gamma = 0.999$. Спецификацию каждого из слоев можно найти в коде работы. Для тренировки использовалось подмножество датасета PTB-XL, состоящее только из примеров, имеющих классы ковалюционной модели из главы 2.3. В качестве тестовой выборки использовался 9 разбиение PTB-XL. В результате тренировки получены метрики из таблицы 2.2

Таблица 2.2 – Результаты валидации модели ecgConvAttentionNet на 9 разбиении PTB-XL на примерах, имеющих только допустимые наборы классов

	precision	recall	f1-score	support
1dAVb	0.67	0.42	0.52	80
RBBB	0.89	0.91	0.90	55
LBBB	0.98	0.81	0.89	54
SB	0.81	0.89	0.85	64
AF	0.84	0.79	0.82	151
ST	0.84	0.78	0.81	83

Для визуализации использовался метод GradCam, каждый входной канал имеет свою карту интереса модели, что позволяет более точно анализировать поведение модели. Однако тот факт, что выход выбранного классификационного канала является выходом стека енкодеров, являющихся сложноинтерпретируемым классификатором, не позволяет интерпретировать работу всего классификатора, но позволяет указать на места интереса первой конволюционной части сети. Интерпретацию работы классификатора можно улучшить, интерпретируя работу второй части сети и агрегируя результаты. На рисунках 2.9- 2.14 изображен результат визуализаций для того же набора примеров, что и в главе 2.4. Благодаря тому, что визуализируемая конволюция имеет большую размерность ядра, чем в случае конволюционной сети (297 против 16), визуализация позволяет точнее определять области интереса классификатора.

Diagnose: AFIB Predicted: ['AFIB']

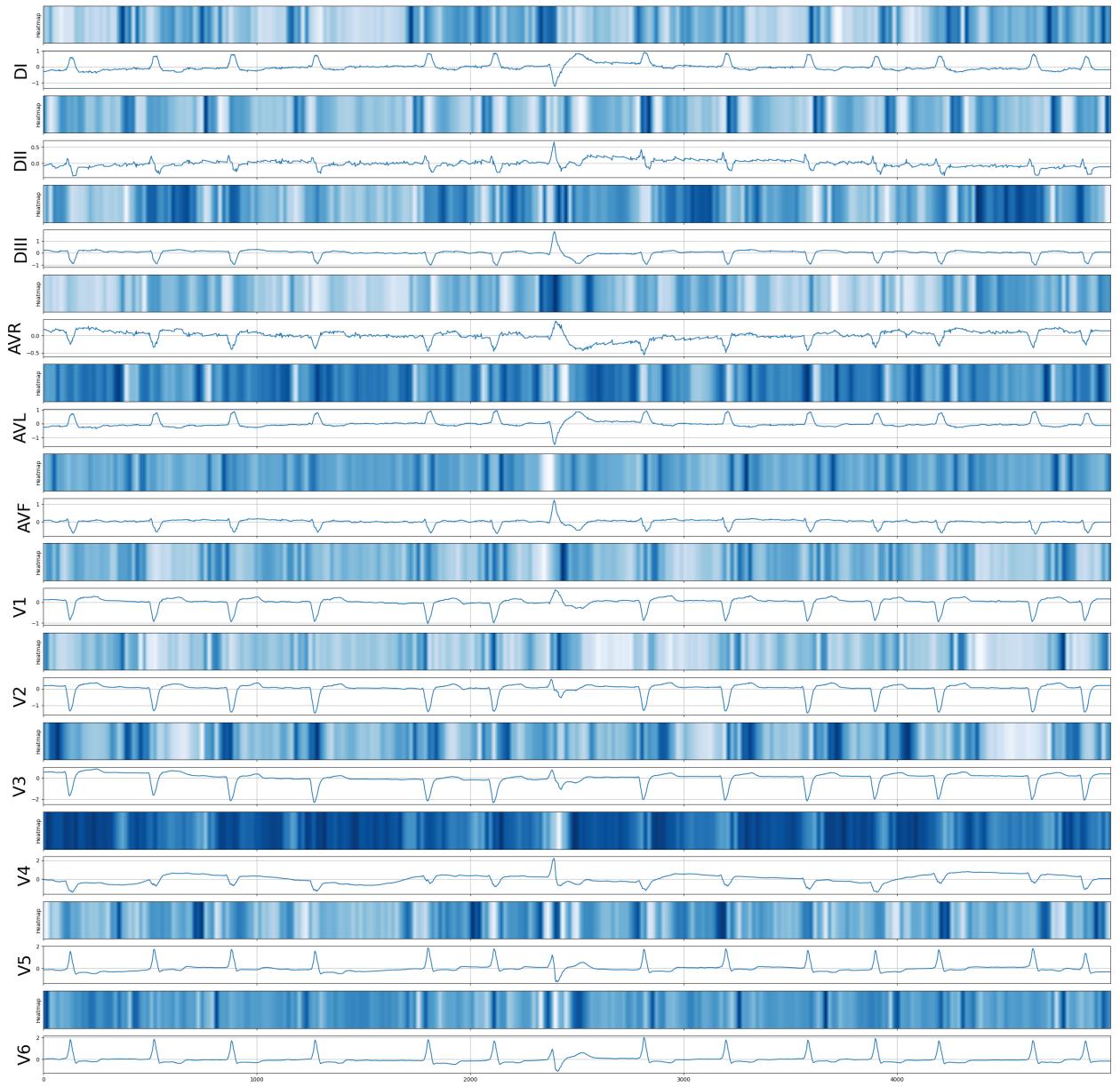


Рисунок 2.9 – Области внимания модели для ЭКГ сигнала с диагнозом

”Мерцательная аритмия”

Diagnose: CRBBB Predicted: ['CRBBB']

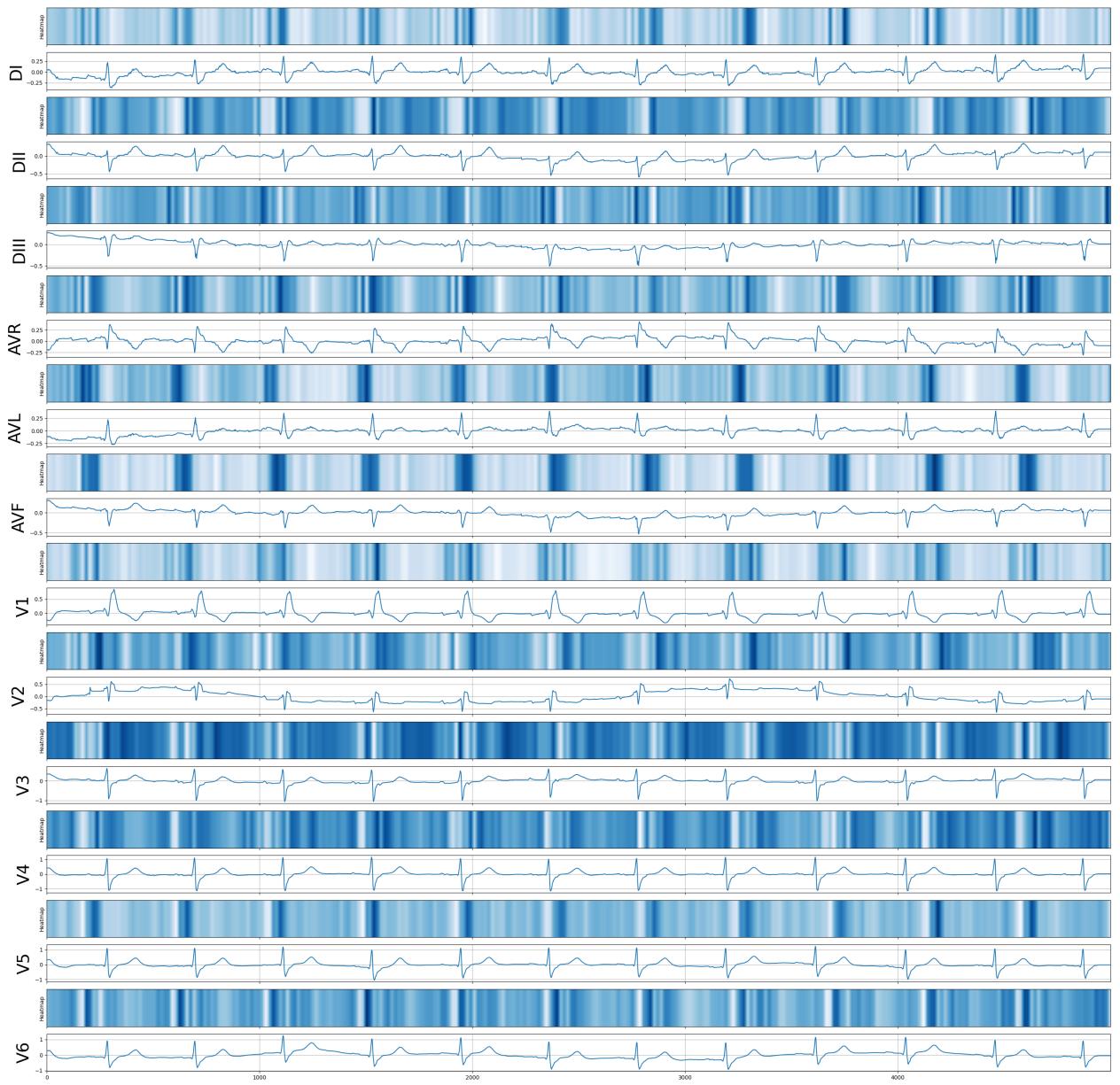


Рисунок 2.10 – Области внимания модели EcgConvAttentionNet для ЭКГ сигнала с диагнозом ”Блокада правой ножки пучка Гиса”

Diagnose: CLBBB Predicted: ['CLBBB']

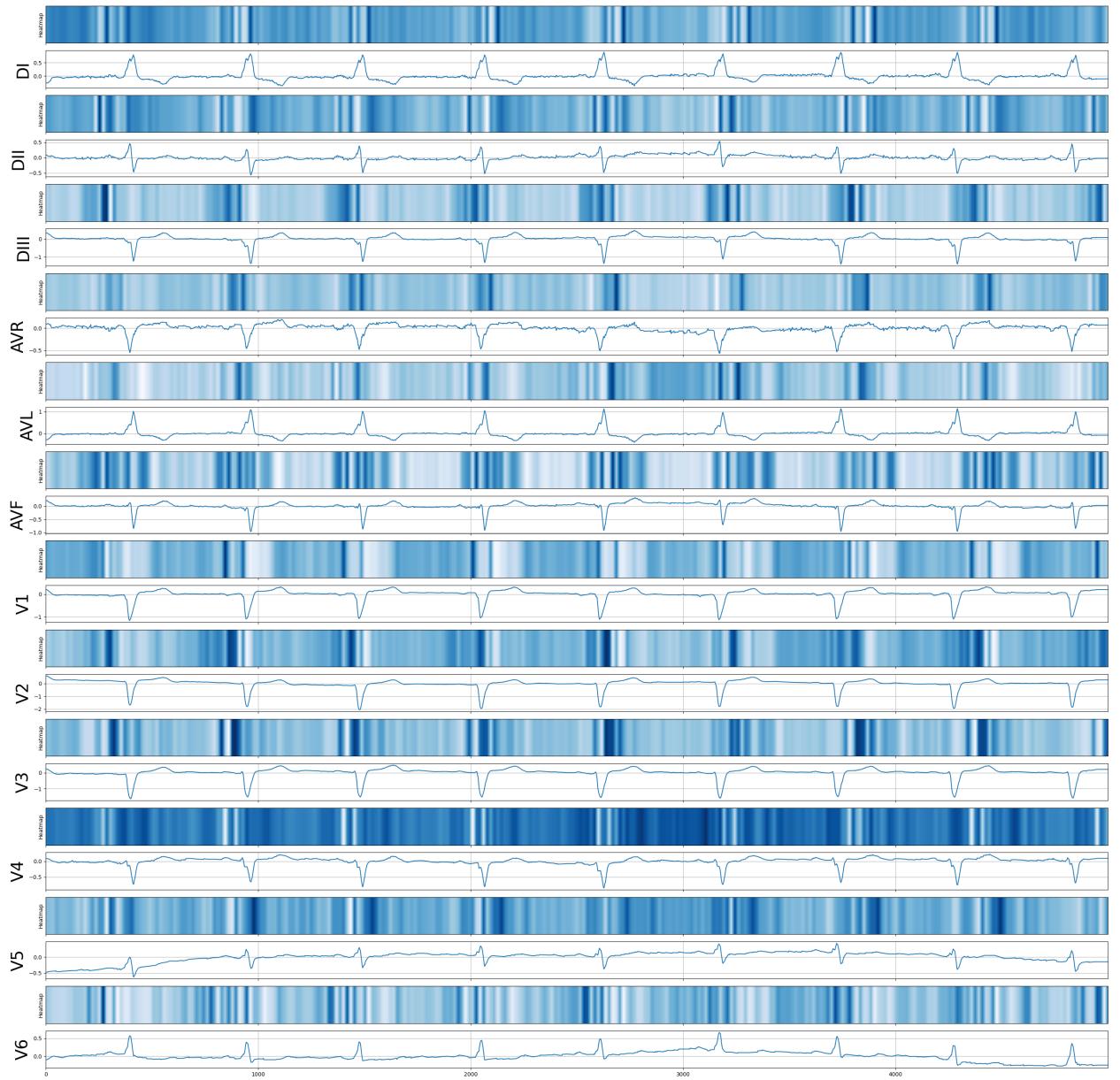


Рисунок 2.11 – Области внимания модели EcgConvAttentionNet для ЭКГ
сигнала с диагнозом ”Блокада левой ножки пучка Гиса”

Diagnose: 1AVB,CLBBB Predicted: ['1AVB', 'CLBBB']

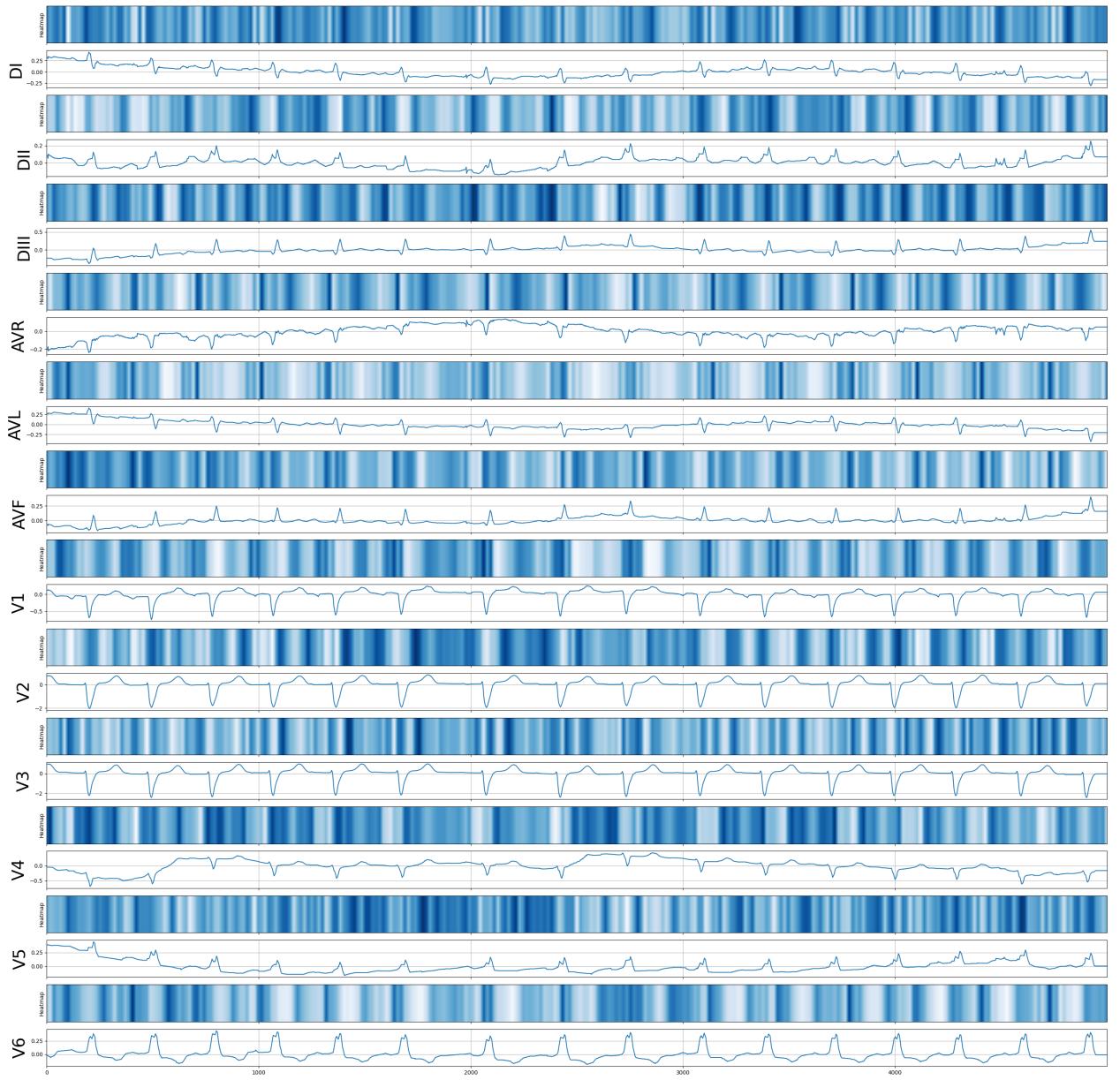


Рисунок 2.12 – Области внимания модели EcgConvAttentionNet для ЭКГ
сигнала с диагнозом ”Атриовентрикулярная блокада первой степени и
блокада левой ножки пучка Гиса ”

Diagnose: SBRAD Predicted: ['SBRAD']

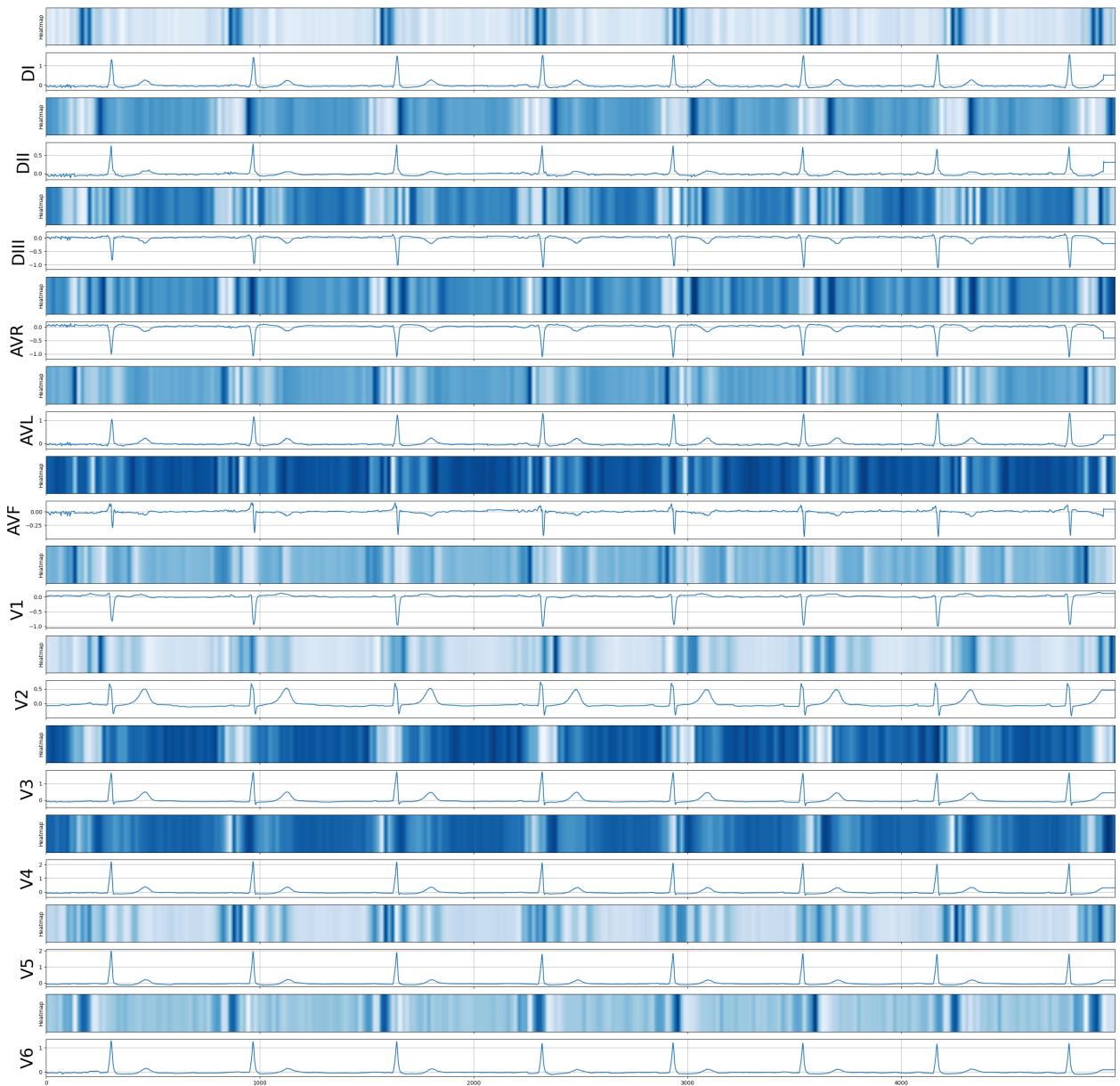


Рисунок 2.13 – Области внимания модели EcgConvAttentionNet для ЭКГ

сигнала с диагнозом ”синусовая брадикардия”

Diagnose: STACH Predicted: ['STACH']

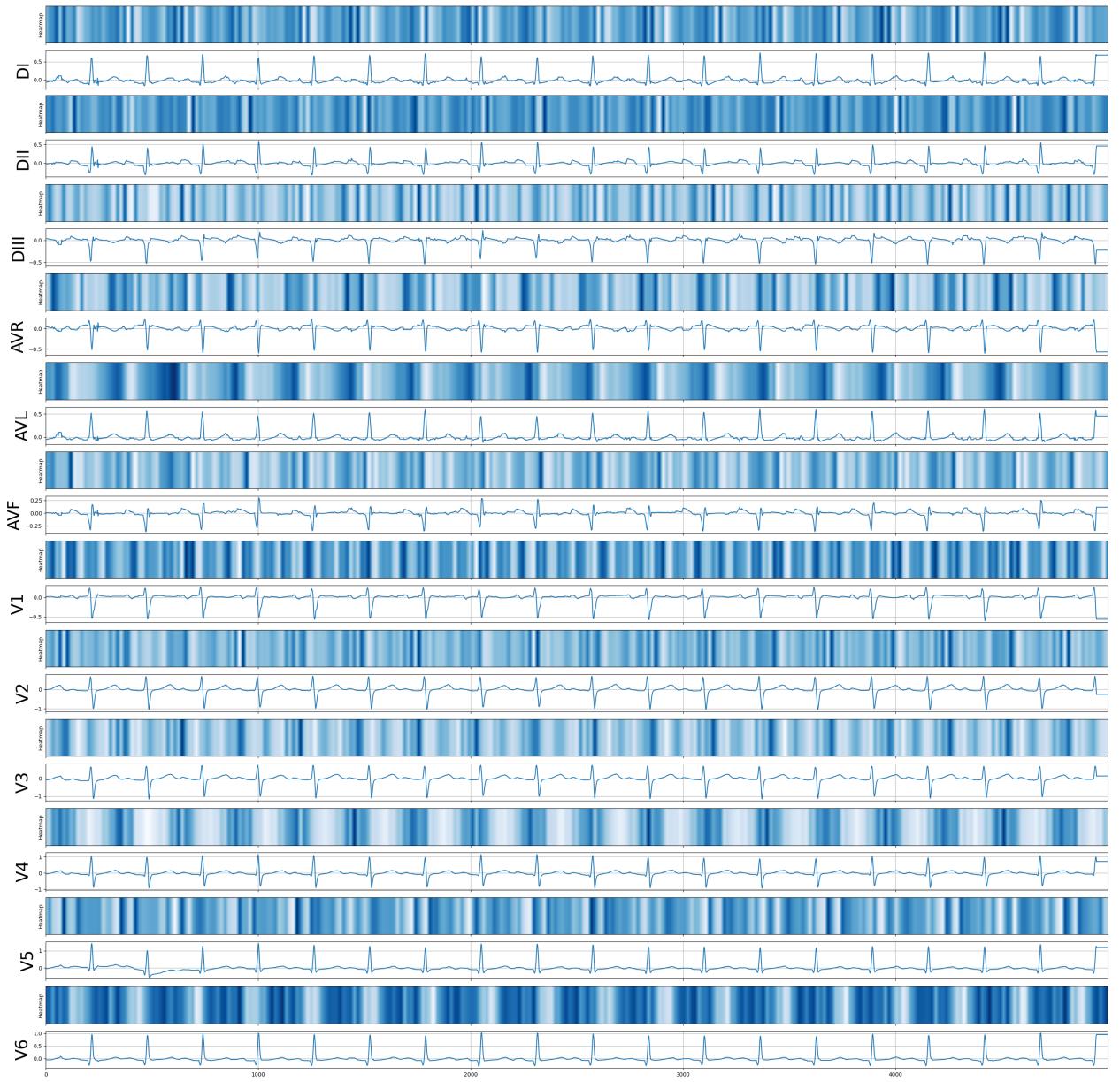


Рисунок 2.14 – Области внимания модели EcgConvAttentionNet для ЭКГ
сигнала с диагнозом ”синусовая тахикардия”

3 Вывод

В ходе работы была представлена и реализована новая классификационная модель ConvAttantionNet, обладающая повышенной интерпретируемой и способностью и конкурентным качеством работы. В продолжение работы возможны архитектурные улучшения сети, такие как увеличение количества каналов свертки и подбор оптимальных значений размеров ядер сверток. Кроме того, для тренировки ecgConvAttentionNet было использовано только подмножество датасета PTB-XL, поэтому возможна тренировка модели на всех имеющихся данных. Кроме того, для подтверждения качества интерпретации необходима консультация прикладного специалиста.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

- 1 Montavon, Grégoire and Samek, Wojciech and Müller, Klaus-Robert: Methods for interpreting and understanding deep neural networks. 2018. Elsevier BV, DOI 10.1016/j.dsp.2017.10.011
- 2 Marco Tulio Ribeiro and Sameer Singh and Carlos Guestrin: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 2016. arXiv 1602.04938
- 3 Jason Yosinski and Jeff Clune and Anh Nguyen and Thomas Fuchs and Hod Lipson: Understanding Neural Networks Through Deep Visualization. 2015. arXiv 1506.06579
- 4 Matthew D Zeiler and Rob Fergus: Visualizing and Understanding Convolutional Networks. 2013. arXiv 1311.2901
- 5 Jost Tobias Springenberg and Alexey Dosovitskiy and Thomas Brox and Martin Riedmiller: Striving for Simplicity: The All Convolutional Net. 2014. arXiv 1412.6806
- 6 Selvaraju, Ramprasaath R. and Cogswell, Michael and Das, Abhishek and Vedantam, Ramakrishna and Parikh, Devi and Batra, Dhruv: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision, 2019, oct p336-359. DOI 10.1007/s11263-019-01228-7
- 7 Ribeiro, A.H., Ribeiro, M.H., Paixão, G.M.M. et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun 11, 1760 (2020). <https://doi.org/10.1038/s41467-020-15432-4>
- 8 Wagner, P., Strodthoff, N., Bousseljot, R., Samek, W., & Schaeffter, T. (2020). PTB-XL, a large publicly available electrocardiography dataset (version 1.0.1). PhysioNet. <https://doi.org/10.13026/x4td-x982>.

9 Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. URL: <https://physionetchallenges.org/2020/>

10 Gonzalez R.C., Woods R.E. Processing digital Image 4 global edition – M: Pearson, 2017. – 993c.

11 Tsung-Yi Lin and Priya Goyal and Ross Girshick and Kaiming He and Piotr Dollár: Focal Loss for Dense Object Detection. 2017. arXiv 1708.02002.

12 Paszke, Adam and Gross, Sam and Massa, Francisco and Lerer, Adam and Bradbury, James and Chanan, Gregory and Killeen, Trevor and Lin, Zeming and Gimelshein, Natalia and Antiga, Luca and Desmaison, Alban and Kopf, Andreas and Yang, Edward and DeVito, Zachary and Raison, Martin and Tejani, Alykhan and Chilamkurthy, Sasank and Steiner, Benoit and Fang, Lu and Bai, Junjie and Chintala, Soumith: PyTorch: An Imperative Style, High-Performance Deep Learning Library. 2019. URL: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>