DOI: https://doi.org/10.30898/1684-1719.2025.12.6

УДК: 621.375.026

МЕТОДЫ РАЗРЕЖИВАНИЯ ДВУНАПРАВЛЕННОЙ LSTM МОДЕЛИ ЦИФРОВОГО КОРРЕКТОРА

Л.И. Аверина, Н.Е. Гутерман, А.А. Хайдаров

Воронежский государственный университет,
394018, Воронеж, Университетская площадь, 1

Статья поступила в редакцию 17 ноября 2025 г.

Аннотация. В работе представлены два подхода к формированию разреженной рекуррентной нейронной сети, используемой в качестве модели цифрового корректора в задаче линеаризации СВЧ усилителя мощности. Проведен сравнительный анализ исследуемых методов с широко распространенным бесструктурным прунингом модели. Изложенная методика позволяет синтезировать цифровой корректор с высоким уровнем разреженности без введения дополнительных гиперпараметров и необходимости тонкой настройки нейронной сети. Разреженная модель, в свою очередь, имеет компактный вид и снижает вычислительную сложность корректора, тем самым допуская эффективное введение цифровых предыскажений в передающем тракте как пользовательского оборудования, так и многоканальных базовых станций.

Ключевые слова: цифровой корректор, нелинейный усилитель мощности, рекуррентная нейронная сеть, разреживающий вариационный дропаут, регуляризация.

Финансирование: Исследование выполнено за счет гранта Российского научного фонда № 24-19-00891, https://rscf.ru/project/24-19-00891/.

Автор для переписки: Гутерман Никита Евгеньевич, n.guterman@internet.ru

Введение

Применение широкополосных сигналов с ортогональным частотным мультиплексированием (OFDM – Orthogonal Frequency-Division Multiplexing) является неотъемлемой частью современных стандартов беспроводной связи. Воздействие такой сложной сигнальной конструкции на СВЧ усилитель мощности передающего тракта приводит как к возникновению искажений в полосе полезного сигнала, так и к росту уровня искажений вне основной полосы [1-3]. Причинами таких негативных эффектов помимо высоких значений отношения пиковой мощности к средней мощности, присущих OFDM сигналам, являются нелинейные и инерционные свойства усилителя. Широко распространенным подходом к устранению искажений на выходе передающего тракта служит цифровая линеаризация [4]. Введение цифровых предыскажений во входное колебание усилителя мощности осуществляется посредством цифрового корректора, построение модели которого определяется архитектурой обучения. В [5] проведен сравнительный анализ основных реализаций, позволяющий выбрать для моделирования в рамках данной статьи архитектуру обучения, основанную на итеративном алгоритме формирования оптимального управления (ILC – Iterative Learning Control).

Общеизвестная практика изложенного подхода к линеаризации побуждает использовать обобщенную линейную модель цифрового корректора. В литературе существует немалое количество таких моделей, представляющих собой простую с точки зрения вычислительной сложности аппроксимацию рядов Вольтерры. Однако представление цифрового корректора в виде сложного нелинейного отображения входного комплексного сигнала кажется наиболее естественным, если использовать в качестве модели искусственную нейронную сеть, эффективность применения которой подтверждается рядом статей, затрагивающих цифровую линеаризацию и поведенческое моделирование усилителей мощности [6-8]. Будучи наделенными свойством универсальной аппроксимации, искусственные нейронные сети избыточны по своей природе, поэтому построение модели с хорошей обобщающей

способностью неразрывно связано с регуляризацией. Цифровая линеаризация усилителя мощности осуществляется с помощью аппаратной реализации, например, на интегральной схеме специального назначения (ASIC – Application-Specific Integrated Circuit). Следовательно, модель цифрового корректора должна быть наименее требовательна в отношении расходуемой памяти и вычислительной мощности. Все вышесказанное стимулирует научное сообщество к поиску методов регуляризации и, как следствие, разреживания искусственных нейронных сетей, используемых в исследуемом вопросе. Тем не менее, большинство встречаемых подходов, нацеленных на сокращение числа параметров модели цифрового корректора, основано на структурном и бесструктурном прунинге [9-13].

В рамках этой статьи для разреживания искусственной нейронной сети, моделирующей цифровой корректор, предложены два метода: разреживающий вариационный дропаут [14], [15] и ℓ_0 –регуляризация [16]. Исходный механизм методов имеет весьма общую структуру и может быть применен к моделям произвольной архитектуры. Для проведения сравнительного анализа была рассмотрена рекуррентная нейронная сеть, представляющая собой двунаправленную LSTM ячейку (LSTM – Long Short-Term Memory) с тремя полносвязными слоями [17]. Особенности реализации исследуемых методов разреживания для рекуррентных моделей подробно представлены в [18-20].

1. Постановка задачи

Задача цифровой линеаризации усилителя мощности подразумевает наличие конкретной архитектуры обучения: размещение цифрового корректора, способ формирования обучающей выборки данных, вид целевой функции. Широко распространенные архитектуры обучения – прямая и обратная (DLA – Direct Learning Architecture, ILA – Indirect Learning Architecture) содержат в постановке задачи линеаризации один характерный недостаток – отсутствие оптимального выходного сигнала для цифрового корректора. Альтернативный подход, базирующийся на эффективном

алгоритме ILC из теории оптимального управления, лишен этого изъяна, устойчив к шумам передающего тракта, а также более удобен для исследовательских целей. Блок-схема архитектуры обучения ILC представлена на рисунке 1.

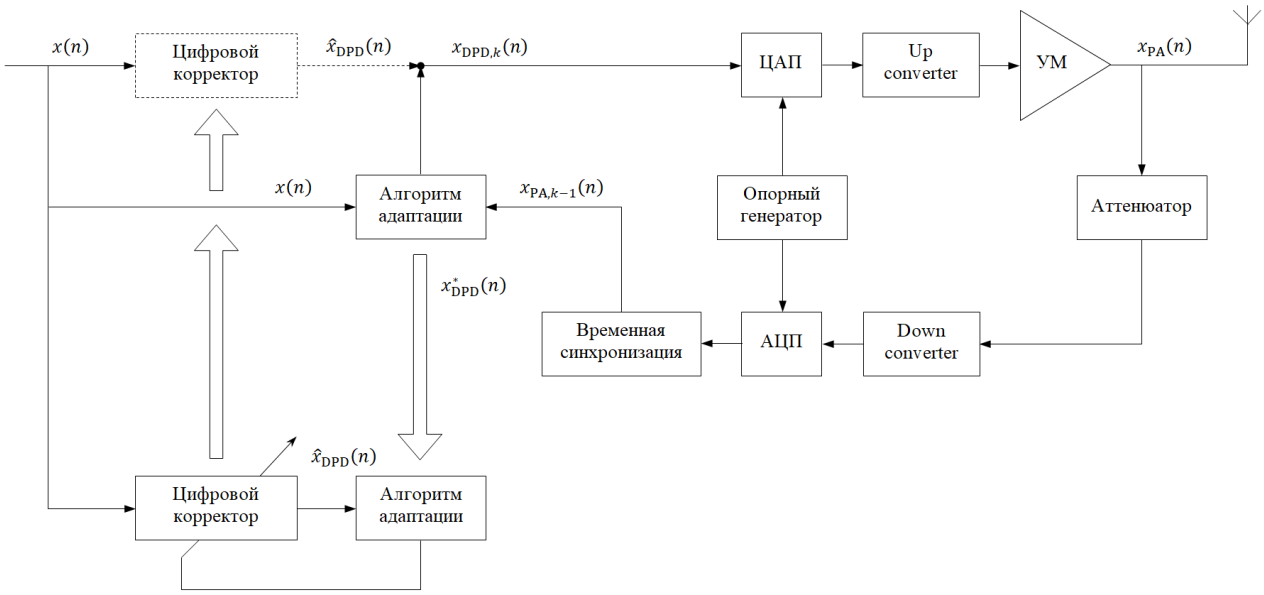

Рис. 1. Блок-схема архитектуры обучения ILC.

Здесь $x(n)$, $x_{\mathrm{DPD}}(n)$, $x_{\mathrm{PA}}(n)$ – оцифрованные в основной полосе частот сигналы на входе передающего тракта, на выходе цифрового корректора и на выходе усилителя мощности соответственно. Суть архитектуры обучения ILC сводится к итеративному алгоритму:

$$\mathbf{x}_{\mathrm{DPD},\,k+1} = \mathbf{x}_{\mathrm{DPD},\,k} + \gamma \mathbf{e}_k\,, \tag{1}$$

где $\mathbf{x}_{\mathrm{DPD}} \in \mathbb{C}^N$ – блочное представление сигнала на выходе цифрового корректора в виде вектора, N – здесь и далее характеризует объем обучающей выборки. Вектор ошибки $\mathbf{e}_k = \mathbf{x}_{\mathrm{PA}} - \hat{\mathbf{x}}_{\mathrm{PA}} \in \mathbb{C}^N$ корректирует входной сигнал усилителя мощности, $\mathbf{x}_{\mathrm{PA}}$ и $\hat{\mathbf{x}}_{\mathrm{PA}}$ – опорный и фактический сигнал на выходе передающего тракта, вещественный параметр γ – скорость обучения. Для линеаризации в качестве $\mathbf{x}_{\mathrm{PA}}$ необходимо выбрать линейно усиленный входной сигнал $G \cdot \mathbf{x} \in \mathbb{C}^N$, где G – коэффициент усиления. Таким образом, в результате последовательных итераций (1) удается сгенерировать

оптимальный выходной сигнал для цифрового корректора $\mathbf{x}^*_{\mathrm{DPD}}$. Цифровая линеаризация далее сводится к классической задаче регрессионного анализа. Необходимо на основе обучающего набора данных $D \equiv \left(\mathbf{x}, \mathbf{x}^*_{\mathrm{DPD}}\right) \equiv \left(\mathbf{x}_i, y_i\right)_{i=1}^{N}$ построить аппроксимирующее параметризованное отображение $F_{\boldsymbol{\theta}} : \mathbb{C}^{N_d} \to \mathbb{C}$, которое и представляет собой модель цифрового корректора.

В качестве модели $F_{\boldsymbol{\theta}}$ в этой статье предлагается выбрать двунаправленную LSTM ячейку с тремя полносвязными слоями, ввиду ее характерных свойств, позволяющих аппроксимировать нелинейную инерционную систему, и схожести с некаузальным представлением в виде рядов Вольтерры [17], [21]. Архитектура рекуррентной модели изображена на рисунке 2. Здесь N_d – длина входной последовательности данных, при этом номер текущего входного комплексного символа в векторе $\mathbf{x}_i$ имеет значение $\lfloor N_d/2 \rfloor$, следовательно, при практической реализации некаузальной модели необходимо ввести задержку в $N_d/2$ символов. Величины N_i^f, N_h^f, N_i^b, N_h^b, N_1, N_2 и N_o описывают размеры слоев: ячейка LSTM для прямого прохождения, аналогично для обратного прохождения и оставшиеся три полносвязных слоя, включая финальное представление квадратур сигнала. Входная последовательность символов $\mathbf{x}_i$ разделяется на действительную $\Re\left(\mathbf{x}_i\right)$ и мнимую $\Im\left(\mathbf{x}_i\right)$ часть, поэтому $N_i^f = N_i^b = N_o = 2$.

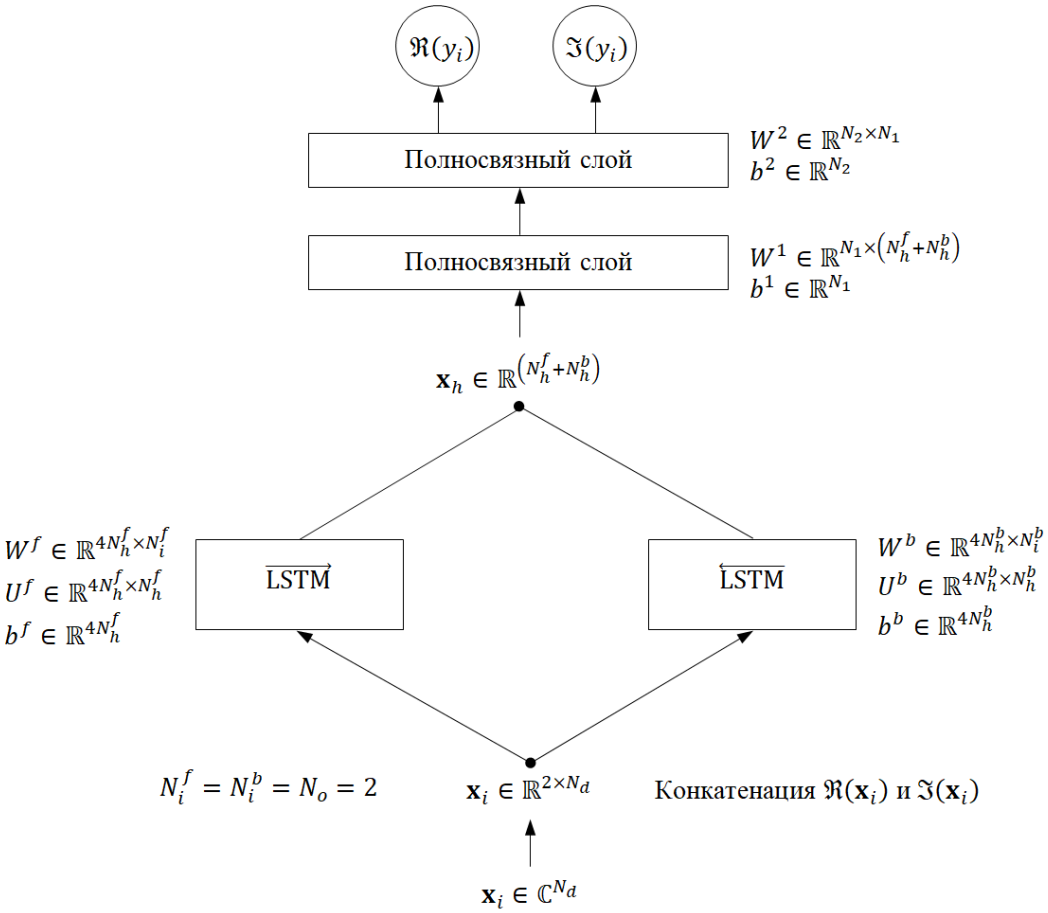

Рис. 2. Архитектура нейронной сети BiLSTM – Bidirectional LSTM.

Разреживание рекуррентной модели, как правило, учитывает только матрицы весовых коэффициентов, сохраняя исходную плотную структуру векторов смещения. Поэтому полный набор параметров имеет вид $\boldsymbol{\theta} = \{\boldsymbol{\theta}_w, \boldsymbol{\theta}_b\}$, где $\boldsymbol{\theta}_w = \{W^f, U^f, W^b, U^b, W^1, W^2, W^3\}$ и $\boldsymbol{\theta}_b = \{b^f, b^b, b^1, b^2, b^3\}$. Неизвестные коэффициенты из $\boldsymbol{\theta}_b$ входят в основное выражение целевой функции для задачи регрессии, в то время как $\boldsymbol{\theta}_w$ также формируют слагаемое регуляризации. Механизм памяти, реализованный в ячейке LSTM, может быть выражен в виде следующих формул:

$$\begin{aligned} i &= \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) & f &= \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \\ g &= \mathrm{th}\left(W_g x_t + U_g h_{t-1} + b_g\right) & o &= \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right), \\ c_t &= f \odot c_{t-1} + i \odot g & h_t &= o \odot \mathrm{th}\left(c_t\right) \end{aligned} \tag{2}$$

где $W_{i,f,g,o} \in \mathbb{R}^{N_h \times N_i}$ – подматрицы $W^{f,b}$, аналогично $U_{i,f,g,o} \in \mathbb{R}^{N_h \times N_h}$ – подматрицы $U^{f,b}$. Выходным сигналом ячейки LSTM является вектор скрытого состояния h_T после прохождения всей последовательности входных данных через модель.

Таким образом, на основе сформированного с помощью ILC обучающего набора данных D необходимо оценить неизвестные параметры рекуррентной модели цифрового корректора. Для этого может быть использован любой из алгоритмов оптимизации нейронных сетей, минимизирующий классическую для задач регрессии целевую функцию. Дальнейшее повествование посвящено методам разреживания, суть которых сводится к регуляризации представленной в этом разделе проблемы.

2. Разреживающий вариационный дропаут

Рассмотрим байесовскую трактовку обучения искусственной нейронной сети. Положим, что исходный набор данных D состоит из статистически независимых одинаково распределенных случайных наблюдений. Необходимо максимизировать параметризованную условную плотность вероятности $p(y_i | \mathbf{x}_i, \boldsymbol{\theta})$, также называемую функцией правдоподобия, относительно множества параметров $\boldsymbol{\theta}$. Байесовский вывод подразумевает наличие некой априорной информации о величинах $\boldsymbol{\theta}$, выражаемой в виде плотности вероятности $p(\boldsymbol{\theta})$. Наблюдаемые данные D позволяют скорректировать исходные предположения относительно $\boldsymbol{\theta}$, представляемые теперь в виде апостериорной плотности вероятности $p(\boldsymbol{\theta} | D)$. Аналитические формулы для расчета функции $p(\boldsymbol{\theta} | D)$ явно по формуле Байеса $p(D | \boldsymbol{\theta}) p(\boldsymbol{\theta}) / p(D)$ крайне сложны с вычислительной точки зрения ввиду комбинаторной природы функции $p(D)$. Выходом из этого положения является вариационный байесовский вывод. Для аппроксимации апостериорной плотности вероятности $p(\boldsymbol{\theta} | D)$ вводится параметризованная функция $q(\boldsymbol{\theta} | \boldsymbol{\alpha})$, степень близости

которой определяется дивергенцией Кульбака-Лейблера $\mathrm{D}_{KL}\left(q(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha})\,\|\,p(\boldsymbol{\theta}\,|\,D)\right)$. Минимизация такого функционала на практике реализуется путем поиска максимума так называемой вариационной нижней оценки обоснованности модели $L(\boldsymbol{\alpha})$:

$$L(\boldsymbol{\alpha}) = -\mathrm{D}_{KL}\left(q(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha})\,\|\,p(\boldsymbol{\theta})\right) + L_D(\boldsymbol{\alpha}), \tag{3}$$

где $L_D(\boldsymbol{\alpha})$ – усредненный по $\boldsymbol{\theta}$ логарифм функции правдоподобия:

$$L_D(\boldsymbol{\alpha}) = \sum_{(\mathbf{x}_i, y_i)\in D} \mathrm{E}_{q(\boldsymbol{\theta}|\boldsymbol{\alpha})}\left[\log p(y_i\,|\,\mathbf{x}_i, \boldsymbol{\theta})\right],$$

$L_D(\boldsymbol{\alpha})$ – искомая целевая функция для обучения модели, а $-\mathrm{D}_{KL}\left(q(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha})\,\|\,p(\boldsymbol{\theta})\right)$ – слагаемое, регуляризирующее оценку параметров $\boldsymbol{\theta}$. Математическое ожидание в выражении (3) не может быть вычислено явно, поэтому для максимизации $L_D(\boldsymbol{\alpha})$ случайная величина $\log p(y_i\,|\,\mathbf{x}_i, \boldsymbol{\theta})$ сэмплируется, что делает возможным применение алгоритмов стохастической оптимизации. Метод Монте-Карло позволяет получить следующую несмещенную дифференцируемую оценку целевой функции [22]:

$$L_D(\boldsymbol{\alpha}) \simeq \hat{L}_D(\boldsymbol{\alpha}) = \frac{N}{M}\sum_{i=1}^{M} \log p\left(y_i\,|\,\mathbf{x}_i, \boldsymbol{\theta} = f(\varepsilon, \boldsymbol{\alpha})\right), \tag{4}$$

где $(\mathbf{x}_i, y_i)_{i=1}^{M}$ – пакет данных из D, $M \le N$. Здесь использован трюк репараметризации, заключающийся в переходе от явного распределения $\boldsymbol{\theta} \sim q(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha})$ к $\boldsymbol{\theta} = f(\varepsilon, \boldsymbol{\alpha})$, где $\varepsilon \sim p(\varepsilon)$ – плотность вероятности шума, а $f(x)$ – дифференцируемая функция. Сэмплирование случайной величины ε вместо $\boldsymbol{\theta}$ нивелирует смещение оценки математического ожидания в (3). Таким образом можно получить несмещенную оценку градиента целевой функции $\nabla_{\boldsymbol{\alpha}}\hat{L}_D(\boldsymbol{\alpha})$.

Вариационный байесовский вывод базируется на выборе в качестве аппроксимации апостериорной плотности вероятности простой с точки зрения дальнейшей оптимизации функции. Так, предполагается, что весовые коэффициенты модели статистически независимы и нормально распределены

$\theta_{ij} = w_{ij} + \sigma_{ij}\varepsilon_{ij}$, а $\varepsilon_{ij} \sim N(0,1)$, следовательно, $q(\theta_{ij} \mid w_{ij}, \sigma_{ij}) = N(w_{ij}, \sigma_{ij}^2)$. Априорная плотность вероятности $\boldsymbol{\theta}$ выбирается таким образом, чтобы исключить параметры θ_{ij} из члена регуляризации $-\mathrm{D}_{KL}(q(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \| p(\boldsymbol{\theta}))$. Единственный закон, удовлетворяющий этому требованию, порождает лог-равномерное распределение вероятности:

$$p(\log|\theta_{ij}|) \propto \mathrm{const} \quad p(|\theta_{ij}|) \propto \frac{1}{|\theta_{ij}|},$$

при этом $p(|\theta_{ij}|)$ – некорректная априорная плотность вероятности. Дивергенция Кульбака-Лейблера в таком случае не может быть выражена явно из-за наличия в выражении математического ожидания $\mathrm{E}_{q(\theta_{ij}|w_{ij},\sigma_{ij})} \log|\varepsilon_{ij}|$, поэтому вместо сэмплирования удобнее прибегнуть к достаточно точной аппроксимации [14]:

$$-\mathrm{D}_{KL}(q(\theta_{ij} \mid w_{ij}, \sigma_{ij}) \| p(|\theta_{ij}|)) \approx$$
$$\approx k_1\sigma\left(k_2 + k_3 \log\frac{\sigma_{ij}}{w_{ij}}\right) - 0.5\log\left(1 + \frac{w_{ij}}{\sigma_{ij}}\right) - k_1, \tag{5}$$
$$k_1 = 0.63576, \quad k_2 = 1.87320, \quad k_3 = 1.48695$$

Таким образом, окончательный механизм обучения представленной рекуррентной нейронной сети с помощью разреживающего вариационного дропаута имеет следующую конструкцию. При прямом прохождении сэмплируется аддитивный гауссовский шум преактиваций $W^{f,b}x_t$ (для каждой последовательности пакета входных данных) и матриц $U^{f,b}$ (единожды для пакета входных данных). Обратное распространение осуществляется по параметрам аппроксимации апостериорной плотности вероятности $\boldsymbol{\alpha} = \{\mathbf{w}, \boldsymbol{\sigma}\}$, множеству математических ожиданий и дисперсий весовых коэффициентов. При этом смещения $\boldsymbol{\theta}_b$ остаются детерминированными. Посредством разреживающей природы дивергенции Кульбака-Лейблера аппроксимация апостериорной плотности вероятности большинства весовых коэффициентов

вырождается в центрированную в нуле $\delta-$образную функцию. Такие параметры могут быть исключены из модели без значительных потерь в точности. Говоря более строго, отбрасыванию подвергаются весовые коэффициенты с низким отношением сигнал/шум $w_{ij}^2/\sigma_{ij}^2<\tau$, τ – пороговое значение. В качестве набора $\boldsymbol{\theta}$ для обученной нейронной сети, как правило, применяют математические ожидания $\mathbf{w}$. Особенности сэмплирования гауссовского шума для преактиваций и весовых матриц рекуррентной модели связаны с локальной репараметризацией, в значительной степени снижающей дисперсию оценки градиента $\nabla_{\boldsymbol{\alpha}}\hat{L}_D(\boldsymbol{\alpha})$ [23].

Изложенный подход порождает бесструктурное разреживание как для ячеек LSTM, так и для полносвязных слоев модели цифрового корректора BiLSTM. Введение групповых весовых коэффициентов, для которых вариационный байесовский вывод имеет идентичную организацию, делает возможным структурное разреживание LSTM ячейки – исключение скрытых нейронов, а также обнуление гейтов, что в свою очередь, делает константными информационные потоки рекуррентной модели [24]. Для того чтобы прояснить сущность написанного, введем в формулы, описывающие логику LSTM, реализации групповых коэффициентов $z^x\in\mathbb{R}^{N_i}$, $z^{i,f,g,o}\in\mathbb{R}^{N_h}$, $z^h\in\mathbb{R}^{N_h}$:

$$\begin{aligned}
i&=\sigma\left(\left(W_i\left(x_t\odot z^x\right)+U_i h_{t-1}\right)\odot z^i+b_i\right)\\
f&=\sigma\left(\left(W_f\left(x_t\odot z^x\right)+U_f h_{t-1}\right)\odot z^f+b_f\right)\\
g&=\mathrm{th}\left(\left(W_g\left(x_t\odot z^x\right)+U_g h_{t-1}\right)\odot z^g+b_g\right)\quad,\\
o&=\sigma\left(\left(W_o\left(x_t\odot z^x\right)+U_o h_{t-1}\right)\odot z^o+b_o\right)\\
c_t&=f\odot c_{t-1}+i\odot g\qquad h_t=o\odot\mathrm{th}\left(c_t\right)\odot z^h
\end{aligned}\tag{6}$$

Соответствующие параметры сэмплируются по аналогии с весовыми коэффициентами $\boldsymbol{\theta}$ для каждой последовательности пакета входных данных. В процессе обучения некоторые групповые весовые коэффициенты также вырождаются в центрированные относительно нуля $\delta-$образные функции. Отбрасывание этих параметров приводит к исключению из матриц LSTM

ячейки строк и столбцов, что не нарушает механизмы памяти LSTM, но значительно повышает эффективность модели с точки зрения вычислительной сложности и особенностей аппаратной реализации. 4 м.

3. ℓ_0 –регуляризация

Целевая функция для обучения искусственной нейронной сети с учетом ℓ_0 –регуляризации имеет вид:

$$L(\boldsymbol{\theta}) = \sum_{(\mathbf{x}_i, y_i) \in D} L_0\left(F_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i\right) + \lambda_0 \|\boldsymbol{\theta}\|_0 \qquad \|\boldsymbol{\theta}\|_0 = \sum_{j=1}^{|\boldsymbol{\theta}|} \mathbf{1}_{\theta_j \neq 0}, \tag{7}$$

где $L_0\left(F_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i\right)$ – функция ошибки присущая конкретной задаче, λ_0 – гиперпараметр регуляризации, $|\boldsymbol{\theta}|$ – мощность множества весовых коэффициентов и $\mathbf{1}_{\theta_j \neq 0}$ – индикаторная функция, обращающаяся в ноль, если $\theta_j = 0$. Такая ℓ_0 –регуляризация способствует уменьшению количества отличных от нуля параметров модели и, как следствие, гарантирует высокий уровень разреженности. К несчастью, оптимизация целевой функции в (7) крайне затруднительна ввиду недифференцируемости нормы ℓ_0 и ее комбинаторной природы, сущность которой проявляется в $2^{|\boldsymbol{\theta}|}$ возможных значениях на множестве весовых коэффициентов.

Введем следующую репараметризацию набора неизвестных параметров $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}} \odot \mathbf{z}, \quad z_j \in \{0,1\}, \quad \tilde{\theta}_j \neq 0, \quad \|\boldsymbol{\theta}\|_0 = \sum_{j=1}^{|\boldsymbol{\theta}|} z_j,$$

где z_j соответствуют бинарным гейтам, которые сигнализируют о наличии или отсутствии весового коэффициента в модели. Полагая $p\left(z_j \mid \pi_j\right) = \mathrm{Bern}\left(\pi_j\right)$, целевая функция, таким образом, содержит член, способствующий росту числа нулевых параметров модели в среднем:

$$L\left(\tilde{\boldsymbol{\theta}},\boldsymbol{\pi}\right)=\mathrm{E}_{p(\mathbf{z}|\boldsymbol{\pi})}\left[\sum_{(\mathbf{x}_i,y_i)\in D}L_0\left(F_{\boldsymbol{\theta}}\left(\mathbf{x}_i\right),y_i\right)\right]+\lambda_0\sum_{j=1}^{|\boldsymbol{\theta}|}\pi_j\,, \tag{8}$$

где случайные величины $\boldsymbol{\theta}$ определяются через $\tilde{\boldsymbol{\theta}}$ и $\mathbf{z}$. Дискретная природа случайных величин $\mathbf{z}$ затрудняет оптимизацию целевой функции $L\left(\tilde{\boldsymbol{\theta}},\boldsymbol{\pi}\right)$. Выходом из этого положения является замена случайной величины Бернулли на близкую по свойствам непрерывную случайную величину, плотность вероятности которой имеет компактный носитель $[0;1]$. Пусть $p(\mathbf{s}\,|\,\phi)$ – параметризованная плотность вероятности случайной величины $\mathbf{s}$. Используя трюк репараметризации для $p(\mathbf{s}\,|\,\phi)$ и метод Монте-Карло для аппроксимации математического ожидания относительно шума $\boldsymbol{\varepsilon}$, получим более удобное с точки зрения оптимизации представление целевой функции:

$$\hat{L}\left(\tilde{\boldsymbol{\theta}},\phi\right)=\frac{N}{M}\sum_{i=1}^{M}L_0\left(F_{\tilde{\boldsymbol{\theta}}}\left(\mathbf{x}_i\right),y_i\right)+\lambda_0\sum_{j=1}^{|\boldsymbol{\theta}|}\left(1-Q\left(s_j\le 0\,|\,\phi\right)\right), \tag{9}$$

где $Q\left(s_j\le 0\,|\,\phi\right)$ – параметризованная функция распределения случайной величины s_j. Бинарные гейты $\mathbf{z}_i$ сэмплируются для каждой последовательности пакета входных данных:

$$\mathbf{z}_i=g\left(f\left(\varepsilon_i,\phi\right)\right),\quad \varepsilon_i\sim N(0,1),\quad g(\mathbf{x})=\min\left(\mathbf{1},\max\left(\mathbf{0},\mathbf{x}\right)\right),$$

где $f(\mathbf{x})$ – детерминированная дифференцируемая функция. В качестве $p(\mathbf{s}\,|\,\phi)$ может быть выбран довольно широкий класс плотностей вероятности. В оригинальном подходе используется следующая схема [16]:

$$u_i\sim U(0,1),\quad s_i=\sigma\left(\left(\log u_i-\log\left(1-u_i\right)+\log\alpha\right)/\beta\right),$$
$$\bar{s}_i=s_i\left(\zeta-\gamma\right)+\gamma,\quad z_i=\min\left(1,\max\left(0,\bar{s}_i\right)\right),$$

где $\phi=\{\alpha,\beta,\gamma,\zeta\}$ – набор гиперпараметров, характеризующих плотность вероятности $p_{\bar{s}}\left(\bar{s}\,|\,\phi\right)$ и функцию распределения $Q_{\bar{s}}\left(\bar{s}\,|\,\phi\right)$ случайной величины $\bar{s}_i$. Следовательно, бинарный гейт z_i распределен непрерывно на интервале

$(0;1)$ и имеет значения $Q_{\bar{s}}(0|\phi)$ и $1-Q_{\bar{s}}(1|\phi)$ в крайних точках промежутка $[0;1]$. Слагаемое регуляризации в (9) выразится так:

$$\sum_{j=1}^{|\boldsymbol{\theta}|}\left(1-Q\left(s_j \le 0|\phi\right)\right)=\sum_{j=1}^{|\boldsymbol{\theta}|}\sigma\left(\log\alpha_j-\beta\log\frac{-\gamma}{\zeta}\right),$$

Для обученной модели в качестве весовых коэффициентов $\boldsymbol{\theta}$ генерируются следующие значения:

$$\hat{\boldsymbol{\theta}}=\tilde{\boldsymbol{\theta}}\odot\hat{\mathbf{z}}\ ,\quad \hat{\mathbf{z}}=\min\left(\mathbf{1},\max\left(\mathbf{0},\sigma\left(\log\boldsymbol{\alpha}\right)\left(\zeta-\gamma\right)+\gamma\right)\right).$$

Здесь стоит отметить, что изложенная схема никак не ограничивает введение в целевую функцию (9) ℓ_1 или ℓ_2-регуляризации. Например, ℓ_2 норму можно представить так:

$$\mathrm{E}_{p(\mathbf{z}|\phi)}\left[\left\|\hat{\boldsymbol{\theta}}\right\|_2^2\right]=\sum_{j=1}^{|\boldsymbol{\theta}|}\left(1-Q_{\bar{s}_j}\left(0|\phi_j\right)\right)\tilde{\theta}_j^2\ .$$

Дальнейший алгоритм оптимизации целевой функции (9) схож с представленным для разреживающего вариационного дропаута. Бинарные гейты, обратившиеся в ноль в процессе обучения модели, могут быть отброшены без существенных потерь в точности. Аналогичным образом проводится и бесструктурное разреживание нейронной сети.

4. Результаты моделирования

Для начала опишем условия генерации обучающего набора данных D. В качестве тестовой последовательности был сформирован OFDM сигнал с количеством поднесущих 1024, шириной полосы 200 МГц и частотой дискретизации $491{,}52$ МГц. Объем обучающей выборки – $24\cdot10^3$, тестовой – $6\cdot10^3$, в число отчетов также входит и циклический префикс OFDM символов. Для реализации ILC количество эпох выбрано равным 120, параметр $\gamma=0{,}07$.

Модель BiLSTM в исходной формулировке содержит 850352 весовых коэффициентов [17], что крайне избыточно и приводит к неоправданным затратам энергии при проведении моделирования. В этой статье размеры слоев

выбраны следующими: $N_h^f = N_h^b = 32$, $N_1 = 64$, $N_2 = 32$, что в сумме дает модель BiLSTM с общим числом параметров 15266. Исходная нейронная сеть достаточно точно моделирует цифровой корректор с точки зрения основных метрик: нормированной среднеквадратической ошибки NMSE (NMSE – Normalized Mean Squared Error) и соотношения просачивания в смежный канал (ACLR – Adjacent Channel Leakage Ratio). При проведении сравнительного анализа была также реализована обобщенная линейная модель – полиномиальная модель с памятью [25], совокупное число параметров которой равно 1365.

Результаты моделирования отражены в таблице 1. Разреживающий вариационный дропаут представлен в двух реализациях: бесструктурное разреживание, при котором априорные распределения вводятся для каждого весового коэффициента; модификация этого метода, включающая структурное разреживание. Эффективность ℓ_0–регуляризации и прунинга модели с последующей тонкой настройкой продемонстрирована только для бесструктурного разреживания. Альтернативный вариант показал крайне неутешительные результаты с точки зрения уровня разреженности и величин основных метрик NMSE и ACLR, поэтому не был включен в таблицу. Гиперпараметры обучения искусственной нейронной сети: число эпох – 300, размер входного пакета данных – 64, оптимизатор – Adam с набором параметров по умолчанию, стартовый темп обучения – 0,01.

Из приведенной таблицы видно, что наибольший уровень разреженности гарантируется структурным разреживающим вариационным дропаутом – сжатие модели приблизительно в 4 раза. Незначительное различие в уровне разреженности между бесструктурным и структурным подходом может быть вызвано недостаточной аппроксимирующей способностью используемой для решения поставленной задачи модели. Увеличение избыточности нейронной сети, то есть добавление полносвязных слоев или повышение размерности вектора скрытого состояния ячейки LSTM, позволит вскрыть относительное различие в уровне разреженности между представленными подходами.

Бесструктурная ℓ_0 – регуляризация порождает менее разреженную рекуррентную модель, но, тем не менее, оба предложенных метода разреживания превосходят бесструктурный прунинг нейронной сети. Стоит отметить, что показатели NMSE и ACLR, представленные в таблице, соответствуют стадиям обучения и валидации нейронной сети, следовательно, величины ошибки и уровня внеполосных искажений могут быть дополнительно снижены путем поиска оптимального набора гиперпараметров.

Таблица 1. Результаты моделирования.

Метод	Размер модели	NMSE, дБ	$ACLR_L$, дБ	$ACLR_R$, дБ
Без DPD	–	–14,87	–23,70	–22,85
Оптимальный сигнал	–	–60,79	–46,06	–47,00
GMP	1365	–27,84/–28,86	–32,41/–33,73	–32,18/–32,94
BiLSTM	**15266**	**–31,82/–34,22**	**–37,23/–39,32**	**–33,71/–36,25**
Вариационный дропаут	4141 (27,1 %)	–30,13/–31,83	–35,84/–37,54	–33,03/–34,29
Вариационный дропаут (структ.)	**3761 (24,6 %)**	**–30,10/–31,79**	**–35,49/–37,23**	**–33,03/–34,29**
ℓ_0-регуляризация	8224 (53,9 %)	–31,37/–33,93	–36,85/–38,89	–34,05/–36,12
Бесструкт. прунинг, порог 0,3	10676 (69,9 %)	–29,84/–30,92	–35,83/–36,87	–32,42/–33,40

Рисунок 3 иллюстрирует разреженную структуру квадратных матриц весовых коэффициентов для скрытого состояния LSTM ячейки $U_{i,f,g,o}$ в результате обучения модели с введением структурного разреживающего вариационного дропаута. Черный цвет элементов матрицы символизирует о нулевом значении параметра, голубым цветом выделены строки и столбцы матриц, исключенные посредством структурного разреживания. Представленная визуализация выделяет общий для всех матриц весовых коэффициентов удаляемый столбец, что позволяет убрать из LSTM ячейки целый информационный поток, так как прямое прохождение в таком случае упростится до вычисления активации соответствующего коэффициента смещения.

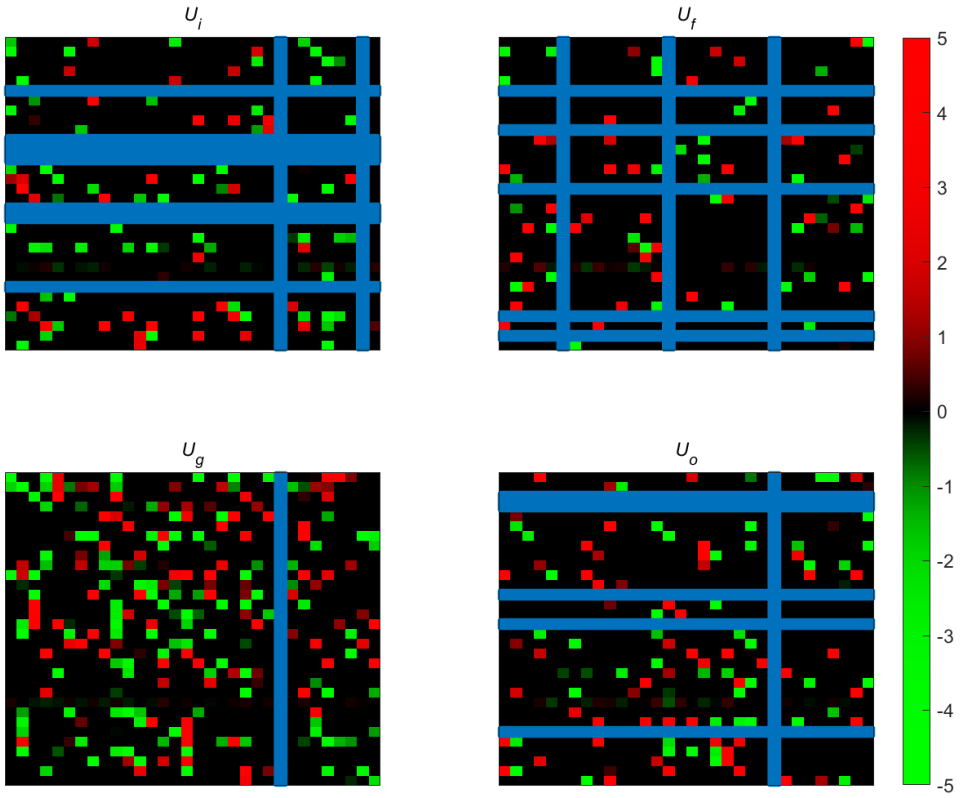

Рис. 3. Разреженная структура матриц LSTM ячейки.

Заключение

Таким образом, в работе представлены возможные методы разреживания рекуррентной нейросетевой модели цифрового корректора для задачи линеаризации СВЧ усилителя мощности передающего тракта. Предложенный разреживающий вариационный дропаут гарантирует наивысший уровень разреженности по сравнению с устоявшимся подходом – прунингом нейронной сети с последующей тонкой настройкой. Эффективный алгоритм, основанный на байесовской трактовке обучения, не нуждается в пороге, определяемом пользователем, и дополнительном этапе тонкой настройки разреженной модели. Регуляризация обучения с введением ℓ_0 нормы демонстрирует меньший уровень разреженности практически без потерь в основных метриках NMSE и ACLR, в то время как вариационный дропаут приводит к незначительному ухудшению NMSE на $\sim 1{,}5$ дБ и ACLR на $\sim 1{,}7$ дБ. Эта негативная тенденция, характерная для методов разреживания, может быть смягчена путем поиска оптимального набора гиперпараметров.

Все вышесказанное позволяет рекомендовать структурный разреживающий вариационный дропаут для формирования эффективных с точки зрения вычислительной мощности и расходуемой памяти моделей цифрового корректора.

Финансирование: Исследование выполнено за счет гранта Российского научного фонда № 24-19-00891, https://rscf.ru/project/24-19-00891/.

Литература

1. Назаров Л.Е., Зудилин А.С. Влияние нелинейностей передатчика на многочастотные сигналы с ортогональным частотным мультиплексированием // Журнал радиоэлектроники. – 2010. – №. 12. – С. 5-5.
2. Аверина Л.И., Шутов В.Д., Рыбалкин Р.А. Бесструктурное моделирование усилителей мощности с учетом инерционных свойств // Известия высших учебных заведений. Радиоэлектроника. – 2013. – Т. 56. – №. 1. – С. 50-57.
3. Ghoniem A.A. et al. An error vector magnitude performance modeling and analysis methodology for 5G mm-Wave Transmitters // IEEE Microwave and Wireless Technology Letters. – 2023. – Т. 33. – №. 6. – С. 771-774.
4. Haider M.F. et al. Predistortion-based linearization for 5G and beyond millimeter-wave transceiver systems: A comprehensive survey // IEEE Communications Surveys & Tutorials. – 2022. – Т. 24. – №. 4. – С. 2029-2072.
5. Аверина Л.И., Гутерман Н.Е. Сравнительный анализ архитектур обучения цифрового корректора в задаче линеаризации широкополосного усилителя мощности // Радиотехника. – 2025. – №. 9. – С. 45-53.
6. Chen P. et al. Behavioral modeling of GaN power amplifiers using long short-term memory networks // 2018 International Workshop on Integrated Nonlinear Microwave and Millimeter-wave Circuits (INMMIC). – IEEE, 2018. – С. 1-3.

7. Аверина Л.И., Бугров О.В. Цифровые корректоры на основе нейронных сетей для линеаризации усилителей мощности // Вестник Воронежского государственного университета. Серия: Физика. Математика. – 2017. – №. 1. – С. 5-14.
8. Xu G. et al. GRU-Attention Model for Linearizing Millimeter-Wave Transmitters in Vehicle to Satellite Communication Systems // IEEE Transactions on Vehicular Technology. – 2024. – Т. 73. – №. 11. – С. 15992-16000.
9. He Z., Tong F. Residual RNN models with pruning for digital predistortion of RF power amplifiers // IEEE Transactions on Vehicular Technology. – 2022. – Т. 71. – №. 9. – С. 9735-9750.
10. Tanio M., Ishii N., Kamiya N. A sparse neural network-based power adaptive DPD design and its hardware implementation // IEEE Access. – 2022. – Т. 10. – С. 114673-114682.
11. Watanabe T., Ohseki T., Kanno I. Hardware-efficient Neural Network Digital Predistortion for Terahertz Power Amplifiers Using DeepShift and Pruning // IEEE Access. – 2025.
12. Li W. et al. GPU-Based Implementation of Pruned Artificial Neural Networks for Digital Predistortion Linearization of Wideband Power Amplifiers // IEEE Journal of Microwaves. – 2025. – Т. 5. – №. 3. – С. 726-738.
13. Liu Z. et al. Low computational complexity digital predistortion based on convolutional neural network for wideband power amplifiers // IEEE Transactions on Circuits and Systems II: Express Briefs. – 2021. – Т. 69. – №. 3. – С. 1702-1706.
14. Molchanov D., Ashukha A., Vetrov D. Variational dropout sparsifies deep neural networks // International conference on machine learning. – PMLR, 2017. – С. 2498-2507.
15. Kingma D.P., Salimans T., Welling M. Variational dropout and the local reparameterization trick // Advances in neural information processing systems. – 2015. – Т. 28.

16. Louizos C., Welling M., Kingma D.P. Learning sparse neural networks through L_0 regularization // arXiv preprint arXiv:1712.01312. – 2017.
17. Sun J. et al. Behavioral modeling and linearization of wideband RF power amplifiers using BiLSTM networks for 5G wireless systems // IEEE Transactions on Vehicular Technology. – 2019. – Т. 68. – №. 11. – С. 10348-10356.
18. Lobacheva E. et al. Structured sparsification of gated recurrent neural networks // Proceedings of the AAAI Conference on Artificial Intelligence. – 2020. – Т. 34. – №. 04. – С. 4989-4996.
19. Lobacheva E., Chirkova N., Vetrov D. Bayesian sparsification of recurrent neural networks // arXiv preprint arXiv:1708.00077. – 2017.
20. Wen L. et al. Structured pruning of recurrent neural networks through neuron selection // Neural Networks. – 2020. – Т. 123. – С. 134-141.
21. Ngoya E., Quindroit C., Nebus J.M. On the continuous-time model for nonlinear-memory modeling of RF power amplifiers // IEEE Transactions on Microwave Theory and Techniques. – 2009. – Т. 57. – №. 12. – С. 3278-3292.
22. Kingma D.P., Welling M. Auto-encoding variational bayes // arXiv preprint arXiv:1312.6114. – 2013.
23. Chirkova N., Lobacheva E., Vetrov D. Bayesian compression for natural language processing // arXiv preprint arXiv:1810.10927. – 2018.
24. Louizos C., Ullrich K., Welling M. Bayesian compression for deep learning // Advances in neural information processing systems. – 2017. – Т. 30.
25. Morgan D.R. et al. A generalized memory polynomial model for digital predistortion of RF power amplifiers // IEEE Transactions on signal processing. – 2006. – Т. 54. – №. 10. – С. 3852-3860.

Для цитирования:

Аверина Л. И., Гутерман Н. Е., Хайдаров А. А. Методы разреживания двунаправленной LSTM модели цифрового корректора // Журнал Радиоэлектроники. – 2025. – №. 12. https://doi.org/10.30898/1684-1719.2025.12.6