МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»

ОТЧЕТ

**Домашнее задание №\_\_1\_\_**

по дисциплине«Методы машинного обучения»

Тема: «Машинный перевод»

ИСПОЛНИТЕЛЬ: \_\_\_\_\_Лю Фэнлинь\_\_\_

ФИО

группа ИУ5И-\_25М\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

подпись

"\_\_"\_\_\_\_\_\_\_\_\_202\_ г.

ПРЕПОДАВАТЕЛЬ: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ФИО

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

подпись

"\_\_"\_\_\_\_\_\_\_\_\_202\_ г.

Москва - 2024

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Задание

Домашнее задание по дисциплине направлено на анализ современных методов машинного обучения и их применение для решения практических задач. Домашнее задание включает три основных этапа:

1. выбор задачи;
2. теоретический этап;
3. практический этап.

Этап выбора задачи предполагает анализ ресурса [paperswithcode.](https://paperswithcode.com/sota) Данный ресурс включает описание нескольких тысяч современных задач в области машинного обучения. Каждое описание задачи содержит ссылки на наиболее современные и актуальные научные статьи, предназначенные для решения задачи (список статей регулярно обновляется авторами ресурса). Каждое описание статьи содержит ссылку на репозиторий с открытым исходным кодом, реализующим представленные в статье эксперименты. На этапе выбора задачи обучающийся выбирает одну из задач машинного обучения, описание которой содержит ссылки на статьи и репозитории с исходным кодом.

## Постановку выбранной задачи машинного обучения, соответствующую этапу выбора задачи.

За последнее десятилетие положение дел в области машинного перевода значительно улучшилось благодаря нейромашинному переводу (НМТ) (Wu et al., 2016), а модели на основе трансформаторов (Vaswani et al., 2017) часто обеспечивают современную производительность перевода (SOTA) при работе с крупными корпорациями (Ott et al., 2018). Наряду с развитием в области НМТ, обучение согласованности (Bachman et al., 2014) получило широкое распространение и показало большие перспективы для повышения эффективности НМТ. Оно просто регуляризирует предсказания модели НМТ, чтобы они были инвариантны к небольшим возмущениям, приложенным к входам (Sato et al., 2019; Shen et al., 2020) и скрытым состояниям (Chen et al., 2021), или к случайности и дисперсии модели, существующей в процедуре обучения (Liang et al., 2021). В частности, Shen et al. (2020) представляют набор методов дополнения данных отсечениями и используют потерю дивергенции Дженсена-Шеннона (JS) для обеспечения согласованности между выходными распределениями исходных и дополненных отсечениями образцов в процедуре обучения. Несмотря на впечатляющие результаты, поиск правильных значений для четырех дополнительных гиперпараметров, введенных при дополнении отсечения, кажется утомительным и трудоемким при наличии ограниченных ресурсов, что снижает его практическую ценность в области НМТ.

Изобретение новых архитектур для нейронного машинного перевода (НМП) имеет фундаментальное значение для прогресса в этой области. От традиционных рекуррентных подходов НМТ перешла к методу самовнимания, который является более эффективным и мощным и установил стандарт для многих других задач НЛП. Другим параллельным направлением исследований является разработка эффективных методов улучшения НМТ без интенсивной модификации архитектуры модели, которые мы будем называть неинтрузивными расширениями. В качестве примера можно привести использование единиц подслова для решения проблемы отсутствия словарного запаса (OOV) или использование дополнительных монолингвальных данных для полусупервизированного обучения с использованием обратного перевода. Одним из основных преимуществ этих методов является применимость к большинству существующих моделей NMT, а также к потенциально будущим архитектурным усовершенствованиям с минимальными изменениями. Таким образом, неинтрузивные расширения используются на практике, чтобы избежать накладных расходов на разработку новых архитектур и расширить возможности существующих современных моделей.

## Теоретический часть отчета

Теоретический этап включает проработку как минимум двух статей, относящихся к выбранной задаче. Результаты проработки обучающийся излагает в теоретической части отчета по домашнему заданию, которая может включать:

* описание общих подходов к решению задачи;
* конкретные топологии нейронных сетей, нейросетевых ансамблей или других моделей машинного обучения, предназначенных для решения задачи;
* математическое описание, алгоритмы функционирования, особенности обучения используемых для решения задачи нейронных сетей, нейросетевых ансамблей или других моделей машинного обучения;
* описание наборов данных, используемых для обучения моделей;
* оценка качества решения задачи, описание метрик качества и их значений;
* предложения обучающегося по улучшению качества решения задачи.

**Часть 1: Bi-SimCut: A Simple Strategy for Boosting Neural Machine Translation**

Мы представляем Bi-SimCut: простую, но эффективную стратегию обучения для повышения производительности нейронного машинного перевода (НМП). Она состоит из двух процедур: двунаправленного предварительного обучения и однонаправленной тонкой настройки. В обеих процедурах используется SimCut - простой метод регуляризации, который обеспечивает согласованность между выходными распределениями исходных и отсеченных пар предложений. Без использования дополнительного набора данных через обратный перевод или интеграции крупномасштабной предварительно обученной модели, Bi-SimCut достигает высокой эффективности перевода в пяти эталонных системах перевода (объем данных варьируется от 160K до 20,2M): Оценка BLEU составила 31,16 для en → de и 38,37 для de → en в наборе данных IWSLT14, 30,78 для en → de и 35,15 для de → en в наборе данных WMT14 и 27,17 для zh → en в наборе данных WMT171 . SimCut - это не новый метод, а версия Cutoff (Shen et al., 2020), упрощенная и адаптированная для NMT, и его можно рассматривать как метод, основанный на возмущении. Учитывая универсальность и простоту SimCut и Bi-SimCut, мы считаем, что они могут служить сильной базой для будущих исследований НМТ.

За последнее десятилетие положение дел в области машинного перевода значительно улучшилось благодаря нейромашинному переводу (НМТ) (Wu et al., 2016), а модели на основе трансформаторов (Vaswani et al., 2017) часто обеспечивают современную производительность перевода (SOTA) при работе с крупными корпорациями (Ott et al., 2018). Наряду с развитием в области НМТ, обучение согласованности (Bachman et al., 2014) получило широкое распространение и показало большие перспективы для повышения эффективности НМТ. Оно просто регуляризирует предсказания модели НМТ, чтобы они были инвариантны к небольшим возмущениям, приложенным к входам (Sato et al., 2019; Shen et al., 2020) и скрытым состояниям (Chen et al., 2021), или к случайности и дисперсии модели, существующей в процедуре обучения (Liang et al., 2021). В частности, Shen et al. (2020) представляют набор методов дополнения данных отсечениями и используют потерю дивергенции Дженсена-Шеннона (JS) для обеспечения согласованности между выходными распределениями исходных и дополненных отсечениями образцов в процедуре обучения. Несмотря на впечатляющие результаты, поиск правильных значений для четырех дополнительных гиперпараметров, введенных при дополнении отсечения, кажется утомительным и трудоемким при наличии ограниченных ресурсов, что снижает его практическую ценность в области НМТ.

В частности, Shen и др. (2020) представили набор методов дополнения данных отсечением и использовали потерю дивергенции Дженсена-Шеннона (JS) для обеспечения согласованности между выходными распределениями исходных и дополненных отсечением образцов в процедуре обучения. Несмотря на впечатляющую производительность, поиск правильных значений для четырех дополнительных гиперпараметров, введенных в аугментацию отсечения, кажется утомительным и трудоемким при наличии ограниченных ресурсов, что препятствует его практическому применению в области НМТ.

В данной работе наша основная цель - предоставить простую, легко воспроизводимую, но труднопреодолимую стратегию для обучения NMT-моделей. Вдохновившись методами увеличения отсечки (Shen et al., 2020) и виртуальной состязательной регуляризации (Sato et al., 2019) для НМТ, мы сначала представляем простой, но эффективный метод регуляризации под названием SimCut. Технически, SimCut не является новым методом и может рассматриваться как упрощенная версия Token Cutoff, предложенная в Shen et al. (2020). Мы показываем, что двунаправленное обратное распространение в регуляризации Куллбэка-Лейблера (KL) играет ключевую роль в улучшении производительности NMT. Мы также рассматриваем SimCut как метод, основанный на возмущениях, и обсуждаем его устойчивость к шумам. Наконец, мотивируясь двунаправленным обучением (Ding et al., 2021) в НМТ, мы представляем Bi-SimCut, двухэтапную стратегию обучения, состоящую из двунаправленного предварительного обучения и однонаправленной доводки, оснащенную регуляризацией SimCut.

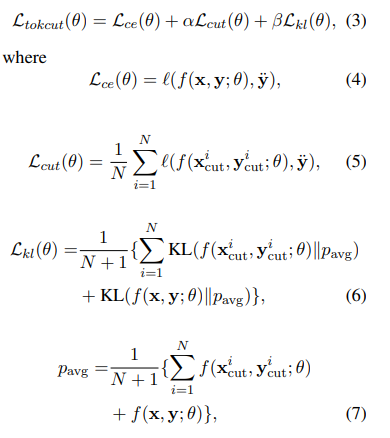
Вклад данной работы можно обобщить следующим образом:

- Мы предлагаем простой, но эффективный метод регуляризации SimCut для улучшения обобщения моделей NMT. SimCut может рассматриваться как метод, основанный на возмущении, и служит сильной базовой основой для подходов робастности. Мы также показываем совместимость SimCut с предварительно обученными языковыми моделями, такими как mBART (Liu et al., 2020).

- Мы предлагаем Bi-SimCut, стратегию обучения для NMT, которая состоит из двунаправленного предварительного обучения и однонаправленной тонкой настройки с регуляризацией SimCut.

- Наши экспериментальные результаты показывают, что обучение НМТ с помощью Bi-SimCut позволяет добиться значительного улучшения по сравнению с моделью Transformer на пяти эталонах перевода (объем данных варьируется от 160K до 20,2M) и превзойти текущий метод SOTA BiBERT (Xu et al., 2021) на нескольких эталонах.

Учитывая пару предложений (x, y), N отсекающих выборок C:\Users\Gong Shengshuo\AppData\Roaming\Tencent\Users\342972224\QQ\WinTemp\RichOle\WE[Q}O`C@)CB4P{R4D69M8Y.png строятся путем случайной установки вкраплений слов x1, ..., xI и y1, ..., yJ равными нулю с вероятностью отсечения pcut. Для каждой пары предложений цель обучения Token Cutoff определяется как:



Настройки. Мы реализуем наш подход поверх Transformer (Vaswani et al., 2017). Мы применяем трансформер с 6 слоями кодера и декодера, 4 головками внимания, размером встраивания 512 и размером слоя FFN 1024. Мы применяем кросс-энтропийные потери с коэффициентом сглаживания меток 0,1 и задаем максимальное количество лексем в партии 4096. Мы используем оптимизатор Adam с бета (0.9, 0.98), 4000 обновлений для разогрева и планировщик скорости обучения с обратным квадратным корнем с начальной скоростью обучения 5e-4 . Мы используем коэффициент отсева 0.3 и декодирование методом лучевого поиска с размером луча 5 и штрафом за длину 1.0. Мы применяем одни и те же конфигурации обучения на этапах предварительного обучения и тонкой настройки, которые будут рассмотрены в следующих разделах. Мы используем файл multi-bleu.pl3 для оценки BLEU (Papineni et al., 2002). Мы обучаем все модели до сходимости на одном графическом процессоре NVIDIA Tesla V100. Все приведенные оценки BLEU относятся к одной модели. Для всех экспериментов, описанных ниже, мы выбираем сохраненное состояние модели с наилучшими показателями валидации.

Bi-SimCut. Мы предлагаем двунаправленное предварительное обучение и однонаправленную тонкую настройку с простой отсекающей регуляризацией (Bi-SimCut), простую, но эффективную стратегию обучения, которая может значительно улучшить обобщение модели NMT. Bi-SimCut состоит из простой отсекающей регуляризации и двухфазной стратегии предварительного обучения и тонкой настройки. Ниже мы представим детали каждой части.

**Часть 2: Data Diversification: A Simple Strategy For Neural Machine Translation**

Мы представляем "Диверсификацию данных" - простую, но эффективную стратегию повышения производительности нейронного машинного перевода (НМП). Она позволяет разнообразить обучающие данные, используя прогнозы нескольких прямых и обратных моделей, а затем объединяя их с исходным набором данных, на котором обучается конечная модель НМТ. Наш метод применим ко всем моделям НМТ. Он не требует дополнительных моноязычных данных, как обратный перевод, и не добавляет дополнительных вычислений и параметров, как ансамбли моделей. Наш метод достигает современных оценок BLEU в 30,7 и 43,7 балла в задачах перевода WMT'14 с английского на немецкий и с английского на французский, соответственно. Он также значительно улучшает результаты в 8 других задачах перевода: 4 задачи IWSLT (англо-немецкий и англо-французский) и 4 задачи перевода с низким уровнем ресурсов (англо-непальский и англо-сингальский). Мы демонстрируем, что наш метод более эффективен, чем дистилляция знаний и двойное обучение, он демонстрирует сильную корреляцию с ансамблями моделей и обменивает недоумение на лучшую оценку BLEU.

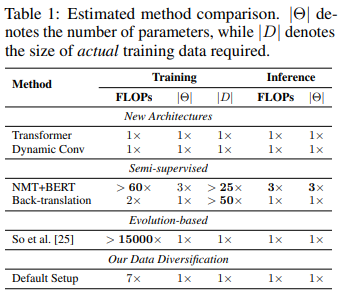
Изобретение новых архитектур для нейронного машинного перевода (НМП) имеет фундаментальное значение для прогресса в этой области. От традиционных рекуррентных подходов НМТ перешла к методу самовнимания, который является более эффективным и мощным и установил стандарт для многих других задач НЛП. Другим параллельным направлением исследований является разработка эффективных методов улучшения НМТ без интенсивной модификации архитектуры модели, которые мы будем называть неинтрузивными расширениями. В качестве примера можно привести использование единиц подслова для решения проблемы отсутствия словарного запаса (OOV) или использование дополнительных монолингвальных данных для полусупервизированного обучения с использованием обратного перевода. Одним из основных преимуществ этих методов является применимость к большинству существующих моделей NMT, а также к потенциально будущим архитектурным усовершенствованиям с минимальными изменениями. Таким образом, неинтрузивные расширения используются на практике, чтобы избежать накладных расходов на разработку новых архитектур и расширить возможности существующих современных моделей.

В этой статье мы предлагаем метод диверсификации данных1 , простой, но эффективный способ последовательного и значительного улучшения машинного перевода. В этом методе мы сначала обучаем несколько моделей на задачах обратного (цель → источник) и прямого (источник → цель) перевода. Затем мы используем эти модели для создания разнообразного набора синтетических учебных данных с обеих языковых сторон, чтобы дополнить исходные данные. Наш подход основан на сочетании нескольких известных стратегий: обратного перевода, ансамбля моделей, увеличения данных и дистилляции знаний для NMT.

Наш метод демонстрирует современное состояние (SOTA) в задачах англо-немецкого и англо-французского перевода WMT'14 с оценками 30,7 и 43,7 BLEU соответственно. Кроме того, он дает выигрыш 1,0-2,0 BLEU в 4 задачах IWSLT (англо-немецкий и англо-французский) и 4 задачах с низким уровнем ресурсов (англо-сингальский и англо-непальский). Мы демонстрируем, что диверсификация данных превосходит другие родственные методы - дистилляцию знаний и двойное обучение, а также дополняет обратный перевод в полусупервизированной установке. Наш анализ также показывает, что метод коррелирует с ансамблями моделей и жертвует сложностью ради лучшего BLEU.

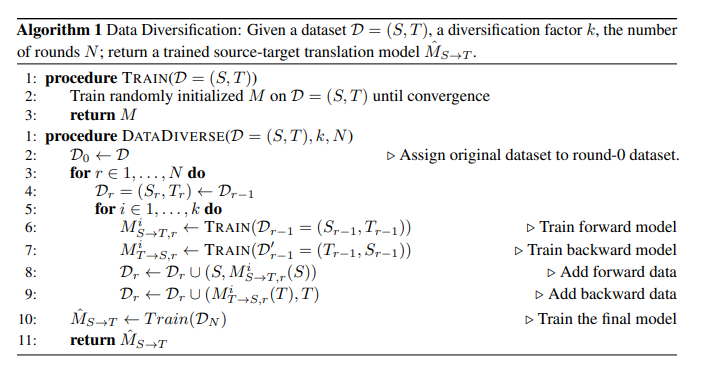
Новые архитектуры. Изобретение новых нейронных архитектур имеет фундаментальное значение для научного прогресса в НМТ. Часто они проходят через дальнейшие усовершенствования и модификации. Например, Шоу и др. и Ахмед и др. предлагают незначительные модификации для улучшения оригинального трансформера с небольшим приростом производительности. Другие предлагают масштабировать процесс обучения на 128 GPU для достижения более значительных улучшений. Ву и др. повторяют цикл с динамической сверткой. Параллельно исследователи ищут и другие дополнительные стратегии для повышения производительности систем NMT, которые ортогональны усовершенствованию архитектур моделей.

Полусамостоятельное НМТ. Полусамостоятельное обучение предоставляет значительные возможности для моделей НМТ. Обратный перевод является простым, но эффективным способом использования дополнительных моноязычных данных. Другой эффективной стратегией является использование предварительно обученных моделей. Zhu и др. недавно предложили новый способ использования предварительно обученных моделей BERT для улучшения NMT. Тем не менее, недостатком обоих подходов является то, что они требуют огромных дополнительных одноязычных данных для обучения/переобучения. Приобретение огромных наборов данных иногда является дорогостоящим, особенно для сценариев с низким уровнем ресурсов (языков или доменов). Более того, в случае использования предварительно обученной BERT, упакованная модель перевода несет дополнительные вычислительные затраты на предварительно обученную модель.



Компромисс между ресурсами. В таблице 1 приведены различные виды затрат на обучение и выводы для различных подходов к улучшению NMT. Разработка новых архитектур, таких как динамическая свертка, практически не дает ощутимых компромиссов при обучении и выводе, но может потребоваться время на доработку и совершенствование новых моделей. С другой стороны, полунаблюдаемые методы часто более просты, но требуют значительно большего количества обучающих данных. В частности, Edunovet al. используют обратный перевод с 50× большим количеством обучающих данных. NMT+BERT требует 60× больше вычислений и 25× больше данных для обучения (включая этап предварительного обучения). Также требуется в 3 раза больше вычислений и параметров во время вывода. Evolved-Transformer, метод, основанный на эволюции, требует более чем в 15 000 раз больше FLOP для обучения. Это может быть нецелесообразно для обычных практиков. С другой стороны, наш метод диверсификации данных прост, как обратный перевод, но не требует дополнительных одноязычных данных. Он также имеет такую же эффективность вывода, как и подход "Новые архитектуры". Однако он вынужден идти на компромисс с дополнительными вычислениями при обучении.

**Метод**



**Связь с существующими методами**

Наш метод имеет определенное сходство с рядом существующих методов, а именно: увеличение данных, обратный перевод, ансамбль моделей, дистилляция знаний и мультиагентное двойное обучение.

Дополнение данных. Наш подход является подлинным методом дополнения данных. Фадаи и др. предложили стратегию дополнения, которая направлена на редкие слова, чтобы улучшить перевод с малым объемом ресурсов. Ванг и др. предложили просто заменять случайные слова другими словами из словарей. Наш подход отличается от этих методов тем, что он не искажает данные случайным образом и не обучает модель на дополненных данных "на лету". Вместо этого он преобразует данные в синтетические переводы, которые следуют различным распределениям модели.

Обратный перевод. Наш метод похож на обратный перевод, который использовался для создания синтетических данных из дополнительных монолингвальных данных целевой стороны. Сеннрих и др. первыми предложили такую стратегию, а Едунов и др. усовершенствовали ее в масштабе. Основное преимущество нашего метода заключается в том, что он не требует дополнительных монолингвальных данных. Наша методика также отличается от предыдущих работ тем, что в ней дополнительно используется прямой перевод, что, как мы показали, очень важно.

Ансамбль моделей. Использование нескольких моделей для усреднения прогнозов и уменьшения дисперсии является типичной особенностью ансамблевых методов. Однако недостатком является то, что параметров тестирования и вычислений требуется в несколько раз больше, чем для отдельной модели. Хотя наш подход к диверсификации коррелирует с ансамблями моделей, он лишен этого недостатка.

Дистилляция знаний. Дистилляция знаний предполагает предварительное обучение большой модели учителя и использование ее предсказаний (прямая трансляция) для обучения меньшей модели ученика в качестве окончательной. По сравнению с этим, наш метод дополнительно использует обратный перевод и задействует несколько прямых и обратных "учителей". Мы используем все данные обратного, прямого перевода, а также исходные данные для обучения финальной модели без какого-либо уменьшения параметров. Мы также повторяем этот процесс несколько раз. В этом контексте наш метод также отличается от метода дистилляции знаний ансамбля, который использует учителей для совместной генерации одной версии данных. Наш метод, напротив, использует преподавателей для индивидуальной генерации различных версий синтетических данных.

Мультиагентное двойное обучение. Мультиагентное двойное обучение предполагает использование двойственности с несколькими агентами прямого и обратного хода. Этот метод объединяет несколько агентов в ансамбль для формирования прямого (Fα) и обратного (Gβ) учителей. Затем он одновременно оптимизирует потери реконструкции ∆x(x, Gβ(Fα(x))) и ∆y(y, Fα(Gβ(y))) для обучения окончательных дуальных моделей. В результате две модели оказываются связанными и запутанными. С другой стороны, наш метод не объединяет агентов таким образом и не оптимизирует какую-либо цель реконструкции.

## Практический часть отчета

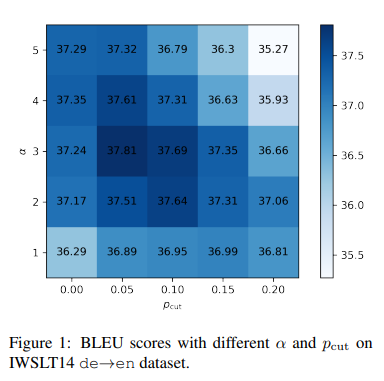
Практический этап включает повторение экспериментов авторов статей на основе представленных авторами репозиториев с исходным кодом и возможное улучшение обучающимися полученных результатов. Результаты проработки обучающийся излагает в практической части отчета по домашнему заданию, которая может включать:

* исходные коды программ, представленные авторами статей, результаты документирования программ обучающимися с использованием диаграмм UML, путем визуализации топологий нейронных сетей и другими способами;
* результаты выполнения программ, вычисление значений для описанных в статьях метрик качества, выводы обучающегося о воспроизводимости экспериментов авторов статей и соответствии практических экспериментов теоретическим материалам статей;
* предложения обучающегося по возможным улучшениям решения задачи, результаты практических экспериментов (исходные коды, документация) по возможному улучшению решения задачи.

**Часть 1: Bi-SimCut: A Simple Strategy for Boosting Neural Machine Translation**

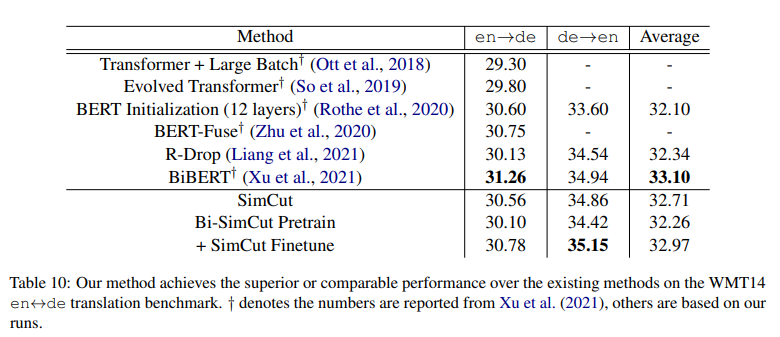
Влияние α и pcut

Здесь мы исследуем влияние скалярных гиперпараметров α и pcut в SimCut. α - это штрафной параметр, который контролирует силу регуляризации в нашей оптимизационной задаче. pcut контролирует процент отсекающих возмущений в SimCut. Здесь мы варьируем α и pcut в пределах {1, 2, 3, 4, 5} и {0.00, 0.05, 0.10, 0.15, 0.20} соответственно и проводим эксперименты на наборе данных IWSLT14 de→en. Обратите внимание, что SimCut упрощается до R-Drop примерно при pcut = 0.00. Тестовые оценки BLEU представлены на рисунке 1. Проверяя производительность модели при различных комбинациях α и pcut, мы получили следующие наблюдения: 1) Слишком маленькое α (например, 1) не может обеспечить такую же хорошую производительность, как большее α (например, 3), что указывает на то, что определенная степень регуляризации при обучении NMT-модели способствует обобщению. В то же время, подавляющая регуляризация (α = 5) не подходит для обучения NMT-моделей. 2) Когда α = 3, наилучшая производительность достигается при pcut = 0,05, а pcut = 0,00 оказывается неоптимальным среди всех выбранных вероятностей. Такое наблюдение показывает, что возмущение отсечения в SimCut может эффективно способствовать обобщению по сравнению с R-Drop.



**Описание набора данных и конфигурация модели**

Для сценария стандартного ресурса мы оцениваем модели NMT на наборе данных WMT14 English-German, который содержит 4,5 млн пар параллельных предложений. Мы объединяем newstest2012 и newstest2013 в качестве валидационного набора и используем newstest2014 в качестве тестового набора. Мы берем предварительно обработанные данные из выпуска5 Сюй и др. (2021), где создан общий словарь с 52K типами BPE. Мы применяем стандартную модель Transformer Big с 6 слоями кодера и декодера, 16 головками внимания, размером встраивания 1024 и размером слоя FFN 4096. Мы применяем кросс-энтропийные потери с коэффициентом сглаживания меток 0.1 и задаем максимальное количество лексем в партии 4096. Мы используем оптимизатор Adam с Beta (0.9, 0.98), 4000 обновлений для разогрева и планировщик скорости обучения с обратным квадратным корнем с начальной скоростью обучения 1e-3 . На этапе тонкой настройки мы уменьшаем скорость обучения до 5e -4. Мы выбираем коэффициент отсева из 0.3, 0.2 и 0.1, основываясь на результатах проверки. Мы используем декодирование методом лучевого поиска с размером луча 4 и штрафом за длину 0,6. Мы обучаем все модели до согласования на 8 графических процессорах NVIDIA Tesla V100. Все приведенные оценки BLEU получены для одной модели.



Результаты

В таблице 10 мы приводим результаты тестирования BLEU всех методов сравнения и нашего подхода на наборе данных WMT14. С помощью двунаправленного предварительного обучения Bi-SimCut и однонаправленной процедуры тонкой настройки наша модель НМТ достигает высоких или SOTA оценок BLEU на эталонах перевода en→de и de→en. В процессе обучения НМТ мы устанавливаем pcut равным 0,05 и настраиваем гиперпараметр α в R-Drop и SimCut на основе результатов на валидационном множестве. Обратите внимание, что показатели BLEU в R-Drop ниже, чем в Liang et al. (2021). Такое расхождение может быть связано с различными этапами предварительной обработки, используемыми в Liang et al. (2021) и Xu et al. (2021). Стоит отметить, что Bi-SimCut превосходит BiBERT по направлению de→en, несмотря на то, что BiBERT включает двунаправленное предварительное обучение, крупномасштабное предварительное обучение контекстуализированных вкраплений и механизм стохастического выбора слоев.

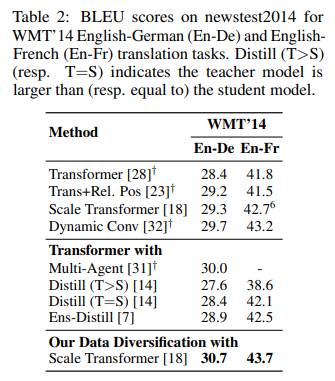
**Часть 2: Data Diversification: A Simple Strategy For Neural Machine Translation**

**Эксперименты**

Мы представляем эксперименты, демонстрирующие, что наш подход к диверсификации данных улучшает качество перевода во многих задачах перевода, включая задачи WMT и IWSLT, а также задачи перевода с высоким и низким уровнем ресурсов. В связи с ограниченным объемом страницы мы кратко описываем постановку каждого эксперимента в соответствующих подразделах и приводим более подробную информацию в Приложении.

**Задачи англо-немецкого и англо-французского перевода WMT'14**

Установка. Мы проводим эксперименты на стандартных задачах перевода WMT'14 с английского на немецкий (En-De) и с английского на французский (En-Fr) языки. Учебные наборы данных содержат около 4,5 и 35 миллионов пар предложений соответственно. Предложения закодированы с помощью кодировки Byte-Pair Encoding (BPE) с 32K операций. Мы используем newstest2013 в качестве набора для разработки и newstest2014 для тестирования. Обе задачи считаются высокоресурсными, поскольку объем параллельных обучающих данных относительно велик. В качестве модели NMT мы используем Transformer и придерживаемся тех же конфигураций, которые были предложены Оттом и другими. При пополнении наборов данных мы отфильтровываем дублирующие пары, в результате чего получаем обучающие наборы данных из 27М и 136М пар для En-De и En-Fr, соответственно.4 Мы не используем никаких дополнительных одноязычных данных.



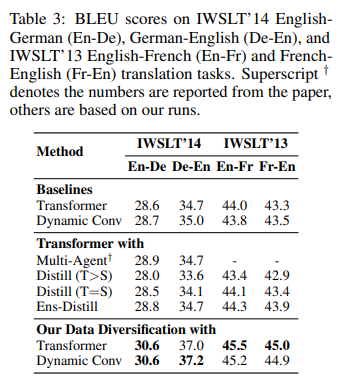
Результаты. Из результатов на тестовом наборе WMT newstest2014 в таблице 2 мы видим, что масштабный трансформатор, который изначально давал 29,3 BLEU в задаче En-De, теперь дает 30,7 BLEU с нашей стратегией диверсификации данных, устанавливая новый SOTA. Наш подход дает улучшение на 1,4 BLEU по сравнению с моделью без диверсификации и на 1,0 BLEU по сравнению с предыдущим SOTA, о котором сообщали Ву и другие. 5 Наш подход также превосходит другие неинтрузивные расширения, такие как мультиагентное двойное обучение и дистилляция знаний, на значительную величину (0,7-3,1 BLEU).

Аналогичное наблюдение можно сделать и для задачи En-Fr WMT'14. Наша стратегия устанавливает новый показатель SOTA в 43,7 BLEU, превышая предыдущий (заявленный) показатель на 0,5 BLEU. Важно отметить, что хотя наш метод увеличивает общее время обучения (включая время обучения базовых моделей), обучение одной модели трансформера за то же время приводит только к переобучению.

**Задачи перевода IWSLT**

Установка. Мы оцениваем наш подход в задачах перевода IWSLT'14 English-German (En-De) и German-English (De-En), IWSLT'13 English-French (En-Fr) и French-English (Fr-En). Учебный набор IWSLT'14 En-De содержит около 160 тысяч пар предложений. Мы произвольно выбираем 5% учебных данных для проверки и объединяем несколько тестовых наборов IWSLT14.TED.{dev2010, dev2012, tst2010, tst1011, tst2012} для тестирования. Набор данных IWSLT'13 En-Fr содержит около 200 тысяч пар предложений для обучения. Мы используем набор IWSLT15.TED.tst2012 для проверки и набор IWSLT15.TED.tst2013 для тестирования. Для всех четырех задач мы используем BPE. Мы сравниваем наш подход с двумя базовыми методами, которые не используют нашу диверсификацию данных: Transformer и Dynamic Convolution.

Результаты. Из таблицы 3 видно, что наш метод существенно и последовательно повышает производительность во всех четырех задачах перевода. В задаче En-De наш метод достигает 30,6 BLEU, что на 2 BLEU выше базового показателя Transformer. Аналогичная тенденция прослеживается и в остальных задачах De-En, En-Fr, Fr-en. Результаты также показывают, что наш метод не зависит от архитектуры модели: и трансформатор, и динамическое преобразование достигают высоких результатов. Напротив, другие методы, такие как дистилляция знаний и мультиагентное двойное обучение, показывают минимальные улучшения в этих задачах.

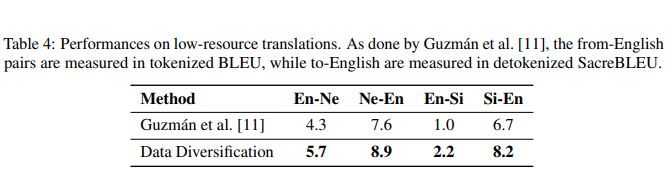


**Задачи перевода с низким уровнем ресурсов**

Продемонстрировав эффективность нашего подхода на языках с высоким уровнем ресурсов, таких как английский, немецкий и французский, мы теперь оценим его работу на языках с низким уровнем ресурсов. Для этого мы используем англо-непальский и англо-сингальский языки с низким уровнем ресурсов, предложенные Гусманом и другими. Непали (Ne) и сингальский (Si) языки являются сложными областями, поскольку источники данных особенно скудны, а словари и грамматики значительно отличаются от языков с большим количеством ресурсов, таких как английский.

Настройка. Мы оцениваем наш метод на контролируемой установке четырех задач перевода с низким уровнем ресурсов: En-Ne, Ne-En, En-Si и Si-En. Мы сравниваем наш подход с базовым в работе Гусмана и др. Параллельные наборы данных English-Nepali и English-Sinhala содержат около 500 тыс. и 400 тыс. пар предложений соответственно. Мы повторяем ту же схему, что и в работе Гусмана и других, и используем их набор dev для разработки и набор devtest для тестирования. В экспериментах по диверсификации данных мы используем k = 3.

Результаты. Из результатов, приведенных в таблице 4, можно заметить, что наш метод последовательно улучшает производительность более чем на 1 BLEU во всех четырех протестированных задачах. В частности, метод достигает 5,7, 8,9, 2,2 и 8,2 BLEU для задач En-Ne, Ne-En, En-Si и Si-En, соответственно. В абсолютном выражении это на 1,4, 1,3, 2,2 и 1,5 BLEU лучше, чем у базовой модели. Без привлечения монолингвальных данных наш метод устанавливает новый уровень техники во всех четырех задачах с низким уровнем ресурсов.



## Выводы обучающегося по результатам выполнения теоретической и практической частей.

В первой статье предлагают Bi-SimCut: простую, но эффективную двухэтапную стратегию обучения для улучшения производительности НМТ. Bi-SimCut состоит из двунаправленного предварительного обучения и однонаправленной процедуры тонкой настройки, оснащенной регуляризацией SimCut для улучшения общности модели НМТ. Эксперименты на эталонах перевода с низким (IWSLT14 en↔de), стандартным (WMT14 en↔de) и высоким (WMT17 zh→en) ресурсом демонстрируют возможности Bi SimCut и SimCut для повышения производительности и устойчивости перевода.

Учитывая универсальность и простоту Bi-SimCut и Sim Cut, мы считаем, что:

1) SimCut можно рассматривать как метод, основанный на возмущениях, и использовать его в качестве сильной основы для исследования робастности.

2) Bi SimCut превосходит многие сложные методы, включающие крупномасштабные предварительно обученные модели или сложные механизмы, и может быть использован в качестве сильной основы для будущих исследований NMT. Мы надеемся, что исследователи возмущений и НМТ смогут использовать SimCut и Bi-SimCut в качестве сильных базовых линий, чтобы сделать полезность и эффективность своих предложенных методов очевидной. В будущей работе мы будем исследовать эффективность SimCut и Bi-SimCut на более последовательных задачах обучения, таких как многоязычный машинный перевод, адаптация домена, классификация текстов, понимание естественного языка и т.д.

Во второй статье предлагают эффективный метод повышения эффективности перевода во многих стандартных задачах машинного перевода. Этот подход достиг наилучших результатов в задаче англо-немецкого перевода WMT'14 с показателем 30,7 BLEU. Он также улучшает результаты в задачах IWSLT'14 "английский - немецкий", "немецкий - английский", IWSLT'13 "английский - французский" и "французский - английский" на 1,0-2,0 BLEU. Кроме того, он превосходит базовые показатели в задачах с низким уровнем ресурсов: английский-непальский, непальский-английский, английский-сингальский, сингальский-английский. Наш экспериментальный анализ показывает, что наш подход демонстрирует сильную корреляцию с ансамблями моделей. Он также обменивает недоумение на лучшую оценку BLEU. Мы также показали, что метод дополняет обратный перевод с дополнительными монолингвальными данными, поскольку он значительно улучшает производительность обратного перевода.

## Список использованных источников

[1] Gao, P., He, Z., Wu, H., & Wang, H. (2022). Bi-simcut: A simple strategy for boosting neural machine translation. *arXiv preprint arXiv:2206.02368*.

[2] Nguyen, X. P., Joty, S., Wu, K., & Aw, A. T. (2020). Data diversification: A simple strategy for neural machine translation. *Advances in Neural Information Processing Systems*, *33*, 10018-10029.

[3] 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), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

[4] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 489–500, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1045. URL https://www.aclweb.org/ anthology/D18-1045.

[5] Sergey Edunov, Myle Ott, Marc’Aurelio Ranzato, and Michael Auli. On the evaluation of machine translation systems trained with back-translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2836–2846, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.253. URL https://www.aclweb.org/anthology/2020.acl-main.253.

[6] Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Data augmentation for low-resource neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 567–573, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2090. URL https://www.aclweb.org/anthology/P17-2090.

[7] Markus Freitag, Yaser Al-Onaizan, and Baskaran Sankaran. Ensemble distillation for neural machine translation. arXiv preprint arXiv:1702.01802, 2017.

[8] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In Jennifer Dy and Andreas Krause, editors, Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 1607–1616, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/furlanello18a.html.