Рецензия на статью “Improving Standard Transformer Models for 3D Point Cloud Understanding with Image Pretraining”

Выполнил: Артем Исмагилов

Суть работы

В статье предлагается модель Pix4Point, основная особенность которой заключается в использовании модели, предобученной на изображениях для обработки облаков точек.

Контекст

Статья была выпущена 25 августа 2022 года исследователями из King Abdullah University of Science and Technology – университета в Саудовской Аравии. Среди авторов есть три PhD студента (Guocheng Qian, Abdullah Hamdi, Xingdi Zhang) и профессор университета (Bernard Ghanem). Все студенты уже несколько лет участвуют в различных публикациях, и имеют около 300-400 цитирований каждый. Профессор имеет около 17000 цитирований и индекс Хирша 55.

Все участники работы до нее занимались компьютерным зрением, в том числе трехмерным. Guocheng Qian, например участвовал в работе PointNeXt, который показывал state of the art результаты на задачах сегментации облаков точек.

Статья достаточно логичным образом вытекает из проблемы моделей, работающих с облаками точек – для этих задач достаточно мало данных, так как собирать и размечать облака точек значительно сложнее, чем собирать и размечать изображения. Для решения этой проблемы логично было попробовать использовать предобучение на обычных изображениях, так как для них есть много данных и способов обучения.

Предшественники и конкуренты

Методы работы с облаками точек стоит разделить на несколько классов:

1. Модели на основе проекции точек в 2D: P2P, MVTN…
2. сверточные и полносвязные модели: PointNet, PointNet++, PointNeXt
3. модели на основе трансформера, модифицированного под облака точек: Point Transformer, PCT
4. модели на основе чистого трансформера: Pix4Point, Point-BERT, Point-MAE

При этом самые хорошие результаты в задачах сегментации и классификации облаков точек показывают модели PointNeXt и Point Transformer, которые и являются основными конкурентами представленной модели.

Исходя из представленных в работе результатов, качество Pix4Point на задаче сегментации облака точек (S3DIS) немного не достигает качества PointNeXt и Point Transformer, показывая mIoU = 69.6 против 70.5 и 70.4 у PointNeXt и Point Transformer соответственно.

При этом в задаче классификации облаков точек (ScanObjectNN), Pix4Point немного превосходит эти модели, имея точность 87.9 против 87.7 и 86.4 у PointNeXt и Point Transformer соответственно.

Однако, авторы статьи не столько сравнивали свою модель с существующими, сколько исследовали эффект от предобучения на изображениях. Предыдущие работы, которые использовали трансформеры без модификаций (Point-BERT, Point-MAE) не смогли достичь результатов, сравнимых с PointNeXt и Point Transformer, проигрывая им порядка 9 процентных пунктов в задачах сегментации и классификации.

В Pix4Point получилось добиться качества, сравнимого с state of the art методами, значительно улучшив качество моделей, основанных на чистом трансформере без модификаций.

Стоит отметить, что для предобучения трансформера в Pix4Point используется метод MAE, но я не считаю что он очень сильно повлиял на результаты, так как не используются какие-то его особенности.

Сильные стороны работы

1. Представленная модель значительно улучшила результаты моделей, основанных на трансформере без модификаций, демонстрируя сравнимое со state of the art качество
2. Использование стандартного трансформера в архитектуре позволяет решать мультимодальные задачи, связанные с 2D изображениями и текстами
3. Проведен ablation study, рассмотрены различные подходы к предобучению трансформера внутри модели, исследовано влияние других элементов на результаты

Слабые стороны работы

1. Возможна мультимодальность стандартного трансформера заявлена как основное преимущество модели, но при этом никак не используется
2. Не исследована зависимость качества от размера взятого трансформера

Дальнейшие исследования

1. Попробовать использовать иерархические трансформеры, например Swin. Я думаю, что такая иерархическая архитектура хорошо подходит для облаков точек. На иерархических архитектурах также основаны другие state of the art методы

2. Эксплуатируя использование в архитектуре стандартного трансформера, решать мультимодальные задачи, например, задачи, возникающие в беспилотных автомобилях, где есть данные с обычных камер и лидаров.
3. Попробовать предобучение модифицированного под облака точек трансформера на обычных изображениях, изменив токенизацию облака точек на токенизацию изображения.