Теоретическое задание №1

Кузина Екатерина

Задание 1

Количество баллов: 2 балла

Прочитайте статью **MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications** и выведите оценку числа параметров для *Depthwise Separable Convolution* слоя. Почему это всё ещё работает, несмотря на очень существенное огрубление тензорных операций?

1. Какую задачу решают авторы? (не более 200 слов)

   Сверточные нейронные сети используются в задачах компьютерного зрения (довольно развивающегося направления), причем общая тенденция заключается в создании более сложных и глубоких сетей для достижения более высокой точности. Основной задачей для авторов этой статьи является построение такой эффективной CNN, чтобы решать поставленные задачи, возможно, не с максимально доступной точностью, но используя меньшие ресурсы для вычисления и не сильно проигрывая в точности. В отличие от других статей, где смотрят на размер нейросети, авторы статьи решили посмотреть на проблему с другой стороны и обратиться к скорости, выраженную в количестве операция mult-adds.

2. Какова основная идея предлагаемого авторами решения поставленной задачи? (не более 300 слов)

   Основная идея решения задачи заключается в следующем: представляемая модель MobileNets основана на depthwise separable convolutions, которая представляят собой разбиение convolution слоя на два: depthwise convolutions и pointwise convolutions. Первая свертка (depthwise convolution) применяет один фильтр для каждого входного канала, а вторая (pointwise convolution (simple 1×1 convolution)) используется для создания линейной комбинации выходных данных предыдущей свертки. После каждого сверточного слоя также применяется нормализация и функция активации ReLU.

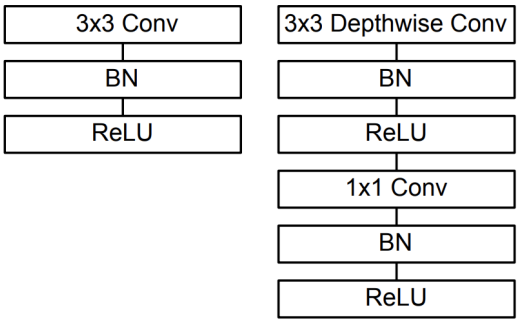

Figure 3. Left: Standard convolutional layer with batchnorm and ReLU. Right: Depthwise Separable convolutions with Depthwise and Pointwise layers followed by batchnorm and ReLU.

На картинке слева представлен обычный сверточный слой с батч нормом и ReLU, справа - введеная авторами depthwise separable convolutions.

Количество вычислений в предложенной свертке меньше в $\frac{1}{N} + \frac{1}{D_k^2}$ раз, где N - число выходных каналов и $D_k * D_k$ - размер ядра первой свертки.

Несмотря на то, что предложенная свертка уже меньше и быстрее по скорости вычисления, авторами были введены 2 гиперпараметра (Width Multiplier/множитель ширины и Resolution Multiplier/множитель расширения) для того, чтобы сделать модель ещё меньше и быстрее.

Смысл множителя ширины α состоит в том, чтобы равномерно прореживать сеть на каждом слое. Для данного слоя и множителя ширины количество входных каналов M становится αM, а количество выходных каналов N становится αN.

Множитель разрешения ρ применяется к входному изображению, и внутреннее представление каждого слоя впоследствии уменьшается на тот же множитель.

3. Каковы результаты, полученные авторами? (оформите в виде списка, не более 200 слов)

   - Использование depthwise separable convolutions по сравнению с обычными свертками снижает точность только на 1% в ImageNet при снижении числа параметров с 29,3 млн до 4,2 млн и уменьшением числа операций с 4866 млн до 569 млн
   - Использование множителя ширины позволяет снизить существенно снизить число операций, однако при этом ухудшение качества логлинейно от количества парамертов итоговой сетки. При уменьшении этого параметра больше чем на 0.5 качество ухудшается намного сильнее.
   - Уменьшение множителя разрешения снижает качество не так быстро, как множитель ширины
   - Исходя из результатов экспериментов применение depthwise separable convolutions может быть использовано во многих задачах компьютерного зрения (распознавания картинок, face attribute classification,

детекции) с не сильным понижением точности, но большим снижением обучаемых параметров.

Оценка числа параметров. Для первого слоя (depthwice convolution) оценка параметров $D_k * D_k * in\_channels$, где $D_k * D_k$ - размер ядра (больше 1*1) и $in\_channels$ - число входных каналов, для второго слоя (pointwise convolutions) оценка числа параметров $in\_channels * out\_channels$, где $out\_channels$ - число выходных каналов. Тогда получаем, что число параметров для данного слоя из двух сверток равно $D_k * D_k * in\_channels + in\_channels * out\_channels$.

Почему это работает. На каждом входном канале фильтры имеют одну и ту же область видимости, когда извлекают признаки. После применяется вторая свертка, которая агрегирует полученные признаки по разным каналам, не сужая область видимости. Получается, что применение depthwise separable convolutions позволяет получать признаки, не сужая область видимости, но огрубая тензорные операции (тк уменьшаются возможные "комбинации"признаков).

Задание 2

Количество баллов: 3 балла

Прочитайте статью **Spherical CNNs** и ответьте на вопросы:

1. Какую задачу решают авторы? (не более 200 слов)

   Сверточные нейронные сети (CNN) уже давно стали основными методами для решения задач компьютерного зрения, например, распознавания и детекции на 2d изображениях. Однако с ростом популярности самоуправляемых автомобилей, беспилотных летательных аппаратов возникла проблема распознавания изображений на сферических изображениях. Сверточные сети позволяют обнаружить локальные закономерности независимо от их расположения на изображении, но они непригодны для распознавания 3D-вращений. Авторы статьи поставили перед собой задачу создать сеть, которая могла бы обнаруживать закономерности независимо от того, как предметы вращаются на сфере, те предлагается вввести математическую модель для сферических CNN.

2. Какова основная идея предлагаемого авторами решения поставленной задачи? (не более 300 слов)

   Из-за того, что преобразования на сфере являются 3D-вращениями, нельзя использовать кросс-корреляци для анализа сферических сигналов, поэтому авторы определяют оператор для вращения фильтров и сферическую кросс-корреляцию между сферическим сигналом и rotated filter, используя введенный оператор. Далее используется обобщенное быстрое преобзование Фурье для подсчета сферических свёрток.

3. Каковы результаты, полученные авторами? (оформите в виде списка, не более 200 слов)

- Вывели теорию для сферических CNN, по факту первыми предложили идею
- Предложили первую автоматически дифференцируемую реализацию обобщенного преобразования Фурье для S^2 (множество точек на сфере с радиусом 1) и $SO(3)$ (набора поворотов в 3d)
- Эмпирически показали полезность сферических CNN для задач обучения, инвариантных к вращению

В чём смысл понятия эквивариантности? Нет, просто формулу не зачтём:) В чём смысл? Являются ли обычные свёртки эквивариантными к каким-либо преобразованиям?

Эквивариантность означает, что можно обнаружить один и тот же объект на повернутом изображении, те: если мы перевели изображение, то сигнал от искомого объекта будет также переведен, но не будет преобразован каким-то другим способом. Обычные свертки являются эквивариантными к преобразованию сдвига и поворота, сохраняющего расстояния.

Зачем нужны и чем отличаются *Spherical Correlation* и *Rotation Group Correlation*?

Если мы хотим решать задачу классификации или регрессии, нам нужно, чтобы голова сети предсказывала числа или логиты. Как это сделать, если внутри сети возникают всякие сферические свёртки?

Постарайтесь объяснить своими словами, как авторы предлагают эффективно вычислять Spherical Correlation или Rotation Group Correlation (достаточно, если объясните одну из них).

Задание 3

Количество баллов: 2 балла

Прочитайте статью **Deep Metric Learning with Spherical Embedding** и опишите метод обучения *Spherical Embeddings* (ответ подкрепите математическими выкладками). Также разберитесь, что такое ablation study и зачем это нужно в принципе и в этой статье в частности.

1. Какую задачу решают авторы? (не более 200 слов)

   DML (Deep Metrinc Learning) в последние годы активно используется в задачах компьютерного зрения (например, распознавания лиц). Важным моментом в DML является разработка функций потерь для оптимизации модели. Предлагается использовать функции потерь таким образом, чтобы расстояние между эмбеддингами двух примеров из одного класса было небольшим, а из разных классов - большим. Например, евклидовое или угловое расстояния могут быть использованы для измерения сходства между двумя объектами, но в функциях потерь DML угловое расстояние используется для определения нормы и направления эмбеддинга, однако норма эмбеддинга играет большую роль в величине градиента (и эта важность игнорируется при обучении модели). Из-за того, что распределение норм эмбеддингов имеет большую дисперсию во время обучения, градиент становится

нестабильным. Это может привести, например, к тому, что обновление направления оптимизации происходит медленее для эмбеддингов с большими нормами. Чтобы устранить данную проблему с градиентом авторы предлагают использовать Spherical Embedding Constraint для лучшей оптимизации нахождения эмбеддингов.

2. Какова основная идея предлагаемого авторами решения поставленной задачи? (не более 300 слов)

   Из-за того, что распределение норм эмбеддингов имеет большую дисперсию, это приводит к тому, что эмбеддингам будет плохо от несбалансированного обновления направления, поэтому авторы статьи предлагаю ограничить норму эмбеддингов во время обучения. Самый простой способ это сделать - ограничить норму эмбеддингов, лежащими на поверхности одной и той же гиперсферы, чтобы они имели одинаковую норму (радиус гиперсферы). Для эмбеддинга, норма которого меньше (больше) средней нормы, SEC пытается увеличить (уменьшить) свою норму при текущем обновлении. Стоит отметить, что данный хак также регулирует общее угловое изменение эмбеддинга: для эмбеддингов с разными нормами угловые изменения будут меньше зависеть от их норм, чем без SEC, что приводит к более сбалансированному обновлению направления эмбеддингов. Также SEC постепенно сокращает разрыв норм на каждой итерации, и по мере того, как дисперсия становится все меньше и меньше, этот негативный эффект еще больше устраняется. Когда дисперсия норм становится небольшой (разные эмбеддинги располагаются почти на одной и той же гиперсфере), величина градиента определяется только угловой зависимостью между эмбеддингами. Авторы статьи также предлагают другой метод регуляризации нормы (регуляризация l2 нормы эмбеддингов, который можно рассматривать как частный случай SEC), но авторы статьи говорят, что при использование SEC модель ведет себя лучше.

3. Каковы результаты, полученные авторами? (оформите в виде списка, не более 200 слов)

   - Проведенные эксперименты показывают, что распределение норм эмбеддингов при использовании SEC для triplet loss более компактно, чем распределение без использования SEC (плюсы от этого описаны в предыдущем пункте). Также для положительных пар распределение их косинусных расстояний становится более компактным, в то время как распределение все еще остается компактным для отрицательных пар. Это указывает на то, что SEC помогает изучить относительно независимую от класса метрику расстояния.
   - При использовании SEC функция потер сходится быстрее (авторы статьи предполагают, что данный эффект достигается именно из-за ограничения на распределение норм эмбеддингов, что в свою очередь обеспечивает более сбалансированное обновление), чем без SEC. Также при увеличении learning rate, скорость сходимости функций потерь с SEC становится быстрее.

Метод обучения *Spherical Embeddings*.
Похожесть между эмбеддингами f_i и f_j:

$$S_{ij} = ||\hat{f}_i - \hat{f}_j||_2^2 \text{ или } S_{ij} =< \hat{f}_i, \hat{f}_j >,$$

где $\hat{f} = \frac{f}{||f||_2}$ - нормализованный эмбеддинг.
В качестве pair-baised loss используется

$$L_{triplet} = (||\hat{f}_a - \hat{f}_p||_2^2 - ||\hat{f}_a - \hat{f}_n||_2^2 + m)_+$$

$$\text{или } L_{tuplet} = \log\left[1 + \sum_n e^{s(<\hat{f}_a, \hat{f}_n> - <\hat{f}_a, \hat{f}_p>)}\right],$$

где m - margin гиперпараметр, s - scale гиперпараметр, $(\hat{f}_a, \hat{f}_n)$ - negative pair (вектора из разных классов), $(\hat{f}_a, \hat{f}_p)$ - positive pair (вектора из одного класса).

Для того, чтобы ограничить норму эмбеддингов используется L_{sec} loss

$$L_{sec} = \frac{1}{N}\sum_{i=1}^{N}(||f_i||_2 - \mu)^2$$

Полная ошибка складывается из ошибки, которую мы хотим минимизировать, и в каком-то смысле члена регуляризации, которым мы ограничиваем норму эмбеддингов: $L = L_{metrica} + \eta L_{sec}$, где $L_{metrica}$, например, $L_{triplet}$ или L_{tuplet}.

Направление эмбеддинга $\hat{f}_t$ находим с помощью стохастического градиентного спуска (или модификации SGD):

$$f_{t+1} = f_t - \alpha\frac{\partial L}{\partial f_t}$$

Осталось найти $\frac{\partial L}{\partial f_t} = \sum\limits_{(i,j)} \frac{\partial L}{\partial S_{ij}} \frac{k}{||f_i||_2}(-\hat{f}_j + cos\theta_{ij}\hat{f}_i) + \frac{2\eta}{N}(||f_t||_2 - \mu)\hat{f}_t$.

η - гиперпараметр, авторы используют $\eta = \sum\limits_{j=1}^{N} ||f_j||_2$.

Что такое ablation study и зачем это нужно. При обучении моделей, часто используют несколько идей, которые могут улучшать общую производительность модели. Чтобы понять, как тот или иной компонент модели влияет на модель в целом, можно оценить модель с выключенным компонентом и измерить её ухудшение. Этот процесс называется ablation study.

В данной статье авторы в экспериментах оценивают, как введеный SEC влияет на модель: они обучают модель с SEC и без него и показывают, какое качество получается при изменении параметров α и η.

Также можно легко вспомнить пример из 1 части курса, когда мы обучали линейные модели с/без регуляризацией и рассматривали то, как работает модель и смотрели на её качество при различных коэффициентах регуляризации.

Задание 4

Количество баллов: 3

Прочитайте статью **LambdaNetworks: Modeling Long-Range Interactions Without Attention** и сравните вычислительную сложность lambda-слоёв со свёрточными слоями и механизмом self-attention (про него можно почитать в учебнике в разделе про нейросети для последовательностей; для картинок суть та же, только смотрим не на токены в последовательности, а на пиксели картинки), а также опишите метод обучения lambda-слоёв (ответ подкрепите математическими выкладками).

1. Какую задачу решают авторы? (не более 200 слов)

   Self-attention является популярным подходом к моделированию долгосрочных зависимостей в данных, но требуемые большие размеры памяти препятствует его применение к длинным последовательностям и многомерным данным, таким как изображения. Linear attention предлагает масштабированное решение для большого использования памяти, но не моделирует внутреннюю структуры данных. Для решения этих проблем авторы предлагают lambda layers, моделирующие долгосрочные взаимодействия между запросом и структурированным набором элементов контекста при сниженных затратах памяти. Также авторы построили нейронные сети на основе lambda layers (Lambda Network), которые являются вычислительно эффективными и моделируют зависимости при небольших затратах памяти и поэтому могут быть применены к большим структурированным входным данным, таким как изображения с высоким разрешением.

2. Какова основная идея предлагаемого авторами решения поставленной задачи? (не более 300 слов)

   Основной идеей lambda layers является построение линейной функции (lambda) для каждого контекста, которая затем применяется к соответствующему запросу. В то время как self-attention определяет сходство между запросом и контекстом, lambda layers суммируют информацию из контекста в линейную функцию фиксированного размера (матрицу), тем самым обходя необходимость в большом использовании памяти.

3. Каковы результаты, полученные авторами? (оформите в виде списка, не более 200 слов)

   - Использование lambda layers позволяет получить качество лучше, чем у моделей с self-attention при меньших параметрах модели
   - lambda layers требует меньше объемов памяти и времени на обучение

Сравните вычислительную сложность lambda-слоёв со свёрточными слоями и механизмом self-attention.

Опишите метод обучения lambda-слоёв.