Distilling the Knowledge in a Neural Network

2022/02/22

守山　慧

著者情報

- Geoffrey Hinton
- Oriol Vinyals
- Jeff Dean

- 会議：NIPS 2014 Deep Learning Workshop

背景

- 音声認識や物体検出タスクでは，巨大なデータセットから特徴を抽出する

- この時使われるモデルは以下のような大きなモデル
  - 複数のモデルを組み合わせたアンサンブルモデル
  - 強い正規化（Dropoutなど）をかけた大きなモデル

背景

- 大きいモデルは実用上以下の問題がある
  - 計算リソースをたくさん必要とする
  - モデルの出力の待ち時間がかかる
  - 学習に時間がかかる

- たくさんの人が使うシステムに活用するのは難しい
  - 結果をなるべく早くユーザーに提供しないといけない
  - 計算リソースが限られている場合が多い

背景

- 大きなモデルから小さなモデルに知識を移す手法を提案
  - 異なるパラメータをもつモデルが同じ入力に対して同じ出力をする状態のこと
  - Rich Caruanaらの論文でアンサンブルモデルから小さなモデルに知識を移すことが可能であることを示した
  - この論文で示された方法は，今回提案する手法の特殊な事例になる

- 新しいアンサンブル手法も提案
  - 学習時間を大幅に短縮する手法

提案手法

- モデルは全てのラベルに対して確率を割り振る
  - ラベル間の確率の大小がある
- この分布にモデルの性能を一般化する情報がある
  - 例：BMWがゴミ収集車と間違えられる確率は低いが，にんじんと間違えられる確率よりも高い
- モデルの最終層に温度付きSoftmax関数を追加
  - Softmax関数に変数Tを加えた関数

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

提案手法

- 確率分布を正解ラベルで修正する方法
  - 今回の実験ではこちらの方法を使う
  - 蒸留に使うデータセットのラベルが既知である場合に有効

- 2つの目的関数の加重平均を蒸留時の目的関数として用いる
  - soft targetに対するCross Entropy
    - 蒸留元のモデルの出力と蒸留先のモデルの出力の分布を合わせる
  - hard targetに対するCross Entropy
    - soft targetの分布を修正する狙いがある ...?

提案手法（既存手法との違い）

- Caruanaらの手法
  - 蒸留元のモデルからの出力を蒸留先モデルの soft targetとして採用
  - soft targetと蒸留先モデルとの二乗和誤差を最小化するように学習する

- この論文で提案する手法
  - 蒸留元のモデルからの出力に温度付き Softmax関数を適用する
  - この出力を蒸留先のモデルの soft targetとして採用している
  - この手法はCrauanaらの手法を一般化したものと等価になる

提案手法(証明)

- 蒸留に使うデータセットの勾配をクロスエントロピーで計算する

$$\frac{\partial C}{\partial z_i} = \frac{1}{T}(q_i - p_i) = \frac{1}{T}\left(\frac{e^{z_i/T}}{\sum_j e^{z_j/T}} - \frac{e^{v_i/T}}{\sum_j e^{v_j/T}}\right)$$

pi:蒸留元モデルのi番目のクラスに対する確率
qi:蒸留先モデルのi番目のクラスに対する確率
zi:蒸留元モデルのi番目のクラスに対する出力
vi:蒸留先モデルのi番目のクラスに対する出力

提案手法(証明)

- 温度Tに対してロジットが十分に小さい時以下のように近似できる

$$\frac{\partial C}{\partial z_i} \approx \frac{1}{T}\left(\frac{1+z_i/T}{N+\sum_j z_j/T} - \frac{1+v_i/T}{N+\sum_j v_j/T}\right)$$

提案手法(証明)

- 各ロジットの平均が0の時下の式が成立するので

$$\sum_j z_j = \sum_j v_j = 0$$

- このように変形できる

$$\frac{\partial C}{\partial z_i} \approx \frac{1}{NT^2}\left(z_i - v_i\right)$$

- ロジットの二乗和誤差を最小化することと等価になる

実験

データセット

- MNIST
  - 60000枚手書き文字の分類

- 音声認識
  - 2000時間の英語発話データ
  - 7億個の音声フレームと対応するHMMの状態確率

実験

MNIST

- 学習に使うモデル
  - 蒸留元：1200個の線形変換ユニットを2層重ねたモデル
  - 蒸留先：800個の線形変換ユニットを2層重ねたモデル

- 実験設定
  - 蒸留先のモデルのテストエラーを比較
    - hard targetで学習
    - 蒸留元の分布（soft target）を使って学習
  - 蒸留時に使うデータのラベルを欠けた状態で学習

実験結果

MNIST

- 蒸留により小さいモデルのテストエラーが大きく改善した
- soft targetは知識の伝達に有効であることがわかる

	蒸留元モデル	蒸留先モデル（蒸留なし）	蒸留先モデル （T=20として蒸留）
テストエラー	67	146	74

実験結果

MNIST

- 学習データからラベル3の画像を抜いて蒸留
  - 蒸留先モデルのテストエラーは 206
  - 133がラベル3の画像によるもの
  - ラベル3のバイアスが低いことが原因
- バイアスを調整した後にスコアを調整
  - テストエラーは109
  - 14がラベル3の画像によるもの
- 良いバイアスの元では98.6%の3の画像を当てることができた

実験

音声認識

- 学習に使うモデル
  - ベースラインモデル：2560個の線形変換ユニットを8層重ねたモデル
  - 蒸留元モデル：ベースラインモデルを 10個アンサンブル
    - ランダムに初期化して，個別に学習したモデルを組み合わせている
  - 蒸留先モデル：ベースラインモデルと同じものを用いる
- データ
  - 26フレームの音声を入力
  - 21番目のHMMの状態確率を予測する

実験結果

音声認識

- 蒸留したモデルはアンサンブルモデルとほとんど同等のパフォーマンス
  - 学習データから使える情報を蒸留により抽出できたと言える

System	Test Frame Accuracy	WER
Baseline	58.9%	10.9%
10xEnsemble	61.1%	10.7%
Distilled Single model	60.8%	10.7%

Table 1: Frame classification accuracy and WER showing that the distilled single model performs about as well as the averaged predictions of 10 models that were used to create the soft targets.

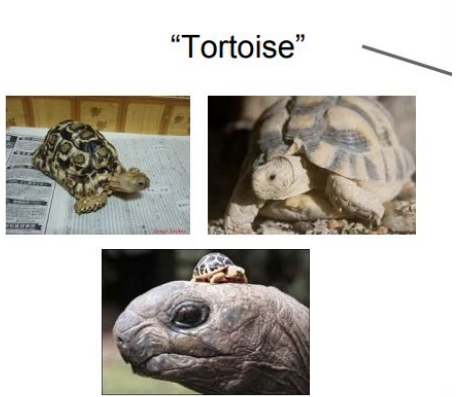
新しいアンサンブル手法の提案

- モデルと学習データが大きい時以下の問題がある
  - 計算リソースが大量に必要
  - 並列化することができない（アンサンブルの利点であるはずなのに！）

- 今回提案するアンサンブル手法
  - ジェネラリストモデルとスペシャリストモデルの2種類のモデルをアンサンブルする
    - ジェネラリストモデル：全てのクラスに対して分類を行うモデル
    - スペシャリストモデル：一部の間違えやすいクラスの分類に特化したモデル

使用するデータセット

- JFTデータセット
  - Googleが収集した15000ラベルの画像データ
  - 1億枚の画像がある
  - ベースラインモデルとしてCNNを学習するのにかかる時間は 6ヶ月



スペシャリストモデル

- ジェネラリストモデルが間違えやすいラベルの分類に特化したモデル
  - 例：違う種類のきのこの分類

- 過学習対策として学習時に以下の工夫をする
  - 重みを学習後のジェネラリストモデルと同じ値で初期化
  - 学習データをしたの方法で分割して学習する
    - 半分を間違えやすいラベル
    - 残りの半分をあまりのデータからランダムにサンプリングする

間違えやすいラベルの決定方法

- 本当のラベルを知らなくても良い方法を用いる

1. ジェネラリストモデルの予測から分散共分散行列を作成
2. この行列に対してK-meansアルゴリズムを用いてm個のグループに分類する

JFT 1: Tea party; Easter; Bridal shower; Baby shower; Easter Bunny; ...
JFT 2: Bridge; Cable-stayed bridge; Suspension bridge; Viaduct; Chimney; ...
JFT 3: Toyota Corolla E100; Opel Signum; Opel Astra; Mazda Familia; ...

Table 2: Example classes from clusters computed by our covariance matrix clustering algorithm

アンサンブルモデルの学習

- 学習方法
  - ジェネラリストモデルの出力を元に n個のクラスを選ぶ
    - 今回の実験ではn=1とする
  - このクラスを担当するスペシャリストモデルを選ぶ(Ak)
  - それらのモデルに対して以下の目的関数を最小化するように学習をする

$$KL(\mathbf{p}^g, \mathbf{q}) + \sum_{m \in A_k} KL(\mathbf{p}^m, \mathbf{q}) \quad (5)$$

KL:カルバック・ライブラー情報量
pg:ジェネラリストモデルが生成した確率分布
pm:全てのスペシャリストモデルが生成した確率分布
Ak:有効になっているスペシャリストモデルの集合
q:T=1とした時のジェネラリストモデルが生成した確率分布

実験結果

- スペシャリストの学習は数日で十分になった
  - ベースラインモデルの学習には 6ヶ月かかっている
- スペシャリストモデルとベースラインモデルのアンサンブルで精度が改善した

System	Conditional Test Accuracy	Test Accuracy
Baseline	43.1%	25.0%
+ 61 Specialist models	45.9%	26.1%

Table 3: Classification accuracy (top 1) on the JFT development set.

実験結果

- 61個のスペシャリストモデル（1つ当たり300ラベル）を作成した
  - 複数のスペシャリストモデルにカバーされるラベルがある
- 選べるスペシャリストモデルの数を変化させて実験

# of specialists covering	# of test examples	delta in top1 correct	relative accuracy change
0	350037	0	0.0%
1	141993	+1421	+3.4%
2	67161	+1572	+7.4%
3	38801	+1124	+8.8%
4	26298	+835	+10.5%
5	16474	+561	+11.1%
6	10682	+362	+11.3%
7	7376	+232	+12.8%
8	4703	+182	+13.6%
9	4706	+208	+16.6%
10 or more	9082	+324	+14.1%

Table 4: Top 1 accuracy improvement by # of specialist models covering correct class on the JFT test set.

正規化としてのsoft target

- この仮説を検証したい
  - soft targetがhard targetに加えてモデルの正規化に必要な情報を多く含む
- 検証方法
  - 音声認識モデルの学習に使うデータの数を変化させる
  - 全体から3%をサンプリングしてhard target，soft targetの両方で学習させる

System & training set	Train Frame Accuracy	Test Frame Accuracy
Baseline (100% of training set)	63.4%	58.9%
Baseline (3% of training set)	67.3%	44.5%
Soft Targets (3% of training set)	65.4%	57.0%

Table 5: Soft targets allow a new model to generalize well from only 3% of the training set. The soft targets are obtained by training on the full training set.

まとめ

- 大きいモデルから小さいモデルに知識を移す一般的な手法を提案
- 大きなデータセットを使って大きなモデルを学習する時の有効な手法を提案
- soft targetにはhard targetよりもモデルの一般化に置いて有効

感想

- 難しかった…

- 理解できたのかできてないのかよくわからない