LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS

- https://arxiv.org/pdf/2106.09685

概要

- fine-tuningの軽量化手法としてLoRA(Low-Rank Adaptation)を提案
- 事前学習されたパラメータを固定して差分を学習することでfine-tuningを実現
- この差分を2つの低ランク行列の積として表現することでfine-tuningする際の学習パラメータ数を減らす

intrinsic dimension

- これからinspiration
- https://arxiv.org/abs/2012.13255
- https://zenn.dev/kabupen/articles/c9f4b5734faeed

LoRA

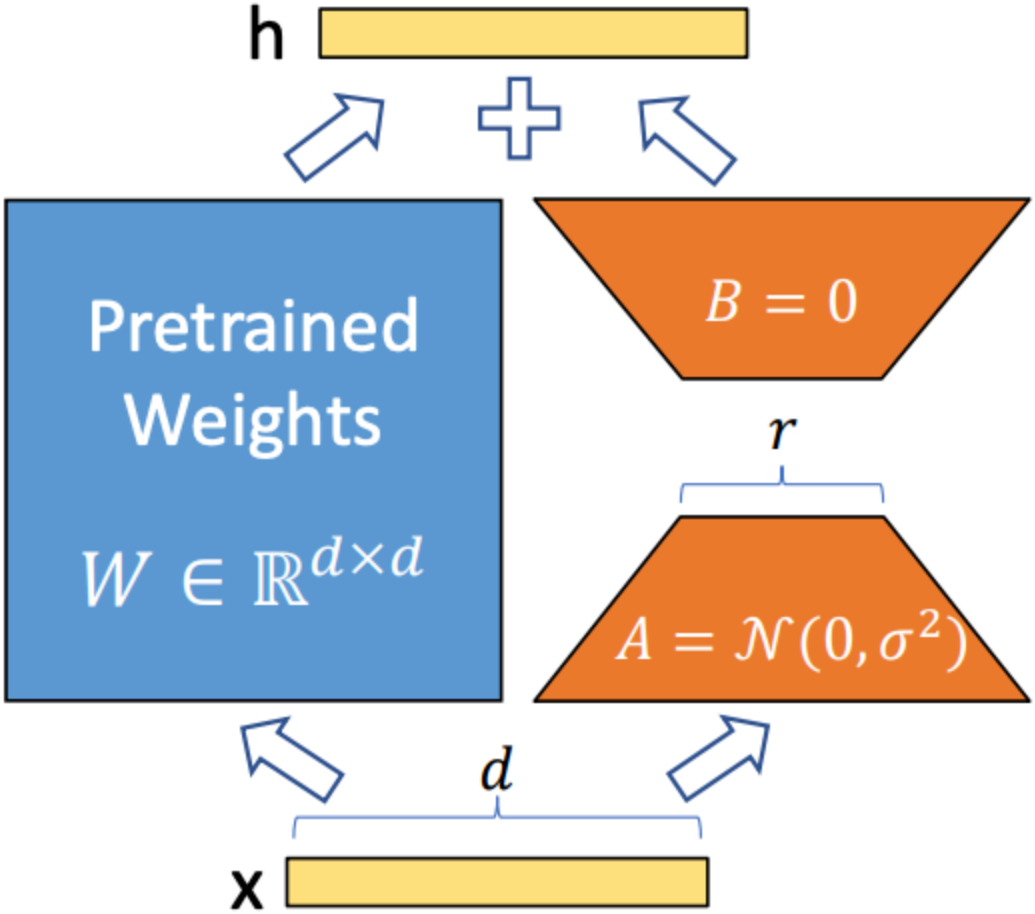
- 事前学習されたconditinal autoregressive language modelを$P_\Phi(y \mid x)$とする
- このときfine-tuningは尤度を最大化することから以下の最適化問題に帰着

$$\max_{\Phi} \sum_{(x,y)\in\mathcal{Z}} \sum_{t=1}^{|y|} \log\left(P_\Phi\left(y_t \mid x, y_{<t}\right)\right)$$

- ここで事前学習された重みをΦ_0としてfine-tuningにより$\Phi_0 + \Delta\Phi$に更新されることを考えてΦ_0を固定して$\Delta\Phi$のみを学習することを考える
- $|\Delta\Phi| = |\Phi_0|$だと意味がないので$\Delta\Phi = \Delta\Phi\left(\Theta\right)$のようにより小さい次元のパラメータ$\Theta$からencodeされる形で考える
- このとき尤度を最大化する問題は以下になる

$$\max_{\Theta} \sum_{(x,y)\in\mathcal{Z}} \sum_{t=1}^{|y|} \log\left(P_{\Phi_0+\Delta\Phi(\Theta)}\left(y_t \mid x, y_{<t}\right)\right)$$

- このように次元を減らしても精度がよいことが実験から示されている
- 例としてGPT-3 175Bのfine-tuningでは$|\Theta|$は$|\Phi_0|$の0.01%まで小さくできる



- LoRAでは2つの低ランク行列$A \in \mathbb{R}^{r \times k}, B \in \mathbb{R}^{d \times r}$を使って$\Delta W \in \mathbb{R}^{d \times k}$を表す
- よって、中間層で$h = W_0 x$のような計算は以下で置き換えられる

$$h = W_0 x + \Delta W x = W_0 x + BAx$$

- fine-tuningする際はAをランダムに正規分布からサンプリングすることで初期化して、Bはゼロ行列で初期化する
- そのため当たり前だけどΔWはゼロ行列から始まる

LoRAの手順

1. **事前学習済みモデルの重みを固定**
   - W_0で固定してfine-tuningで更新しないようにする
2. **低ランク行列の注入**
   - A, Bを初期化してBAで各層の重みの更新を表現
3. **学習**

- A, Bのみを学習

4. **適応と推論**
    - 推論時には新しい重み$W_0 + BA$を使う
    - これにより推論遅延を増加させることなくモデルを適応させられる

英語

- agnostic : 不可知論的な
    - 依存しないという意味もあり、OS-agnosticなどでOSに依存しないという意味になる
- terminology : 用語、専門用語