GPT

静岡大学 情報学部 情報科学科
峰野研究室
B4 原田海斗

GPTとは

GPT (Generative Pretrained Transformer)

○ Transformerベースの学習済み大規模言語モデル

― 教師なし学習と教師あり学習を組み合わせた学習手法(半教師あり学習)

教師なし学習フェーズ

大容量言語モデル学習

$$h_0 = UW_e + W_p$$
$$h_l = transformer\_block(h_{l-1}) \forall i \in [1, n]$$
$$P(u) = softmax(h_n W_e^T)$$

$P(u)$はTransformerデコーダで計算される

$$L_1(U) = \sum_i \log P(u_i | u_{i-k}, \dots, u_{i-1}; \theta)$$

尤度$L_1(U)$を最大化するθを求める
確率的勾配降下法を用いて探索

教師あり学習フェーズ

ファインチューニング

対象データセット$C = \{x^1, \dots, x^m\}$を想定

$$P(y | x^1, \dots, x^m) = softmax(h_l^m W_y)$$

事前学習済みモデルによって出力を得る

$$L_2(C) = \sum_{(x,y)} \log P(y | x^1, \dots, x^m)$$
$$L_3(C) = L_2(C) + \lambda * L_1(C)$$

教師なしフェーズと同等の目的関数$L_2(C)$
学習済みのパラメータも含めて学習する

RADFORD, Alec, et al. Improving language understanding by generative pre-training. 2018.

GPTとは

GPT (Generative Pretrained Transformer)

○ Transformerベースの学習済み大規模言語モデル

― 教師なし学習と教師あり学習を組み合わせた学習手法(半教師あり学習)

教師なし学習フェーズ

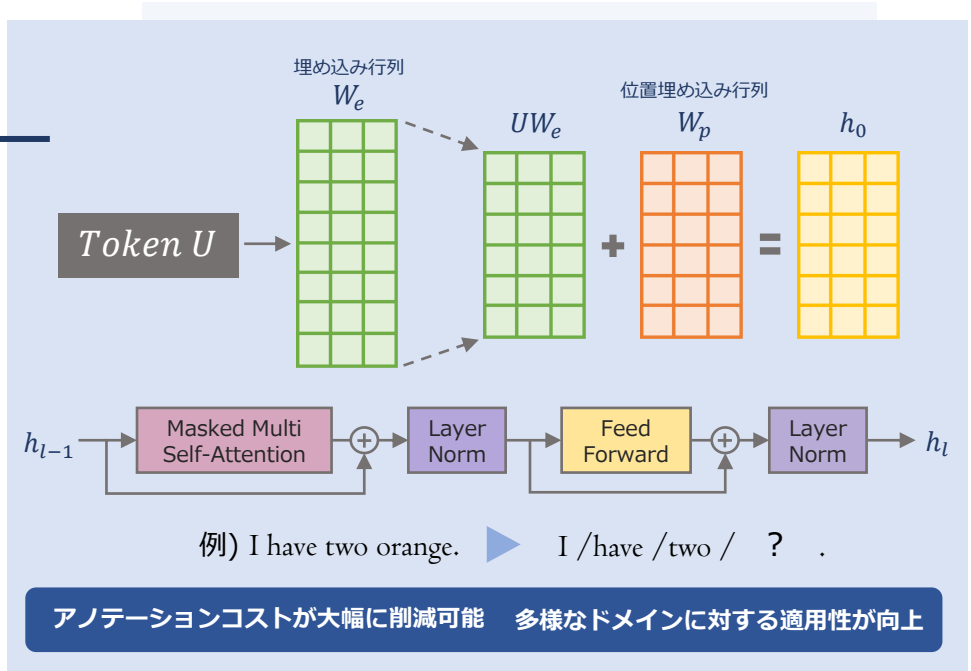
大容量言語モデル学習

$$h_0 = UW_e + W_p$$
$$h_l = transformer\_block(h_{l-1}) \forall i \in [1, n]$$
$$P(u) = softmax(h_n W_e^T)$$

$P(u)$はTransformerデコーダで計算される

$$L_1(U) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \theta)$$

尤度$L_1(U)$を最大化するθを求める
確率的勾配降下法を用いて探索



BANDY, Jack; VINCENT, Nicholas. Addressing" documentation debt" in machine learning: A retrospective datasheet for bookcorpus. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1). 2021.

GPTとは

事前学習用データセットについて

○ 教師なし学習用データセット $U = \{u_0, u_1, \ldots, u_{n-1}, u_n\}\ (n < k)$

― BookCorpus

… 未発表著者による全16ジャンルの無料小説本(11038冊分)に関する大規模テキストコーパス

… 自費出版電子書籍プラットフォーム「Smashwords」から作成され, 公式バージョンは非公開

約7400万行
(約4.5GB)

Text
his platinum blond hair and blue eyes were completely hers.
it was only his build that he was taking after his father.
where megan was a diminutive 5'3", davis was 6'1"and two hundred pounds.
mason was already registering off the charts in height and weight according to his pediatrician.

Smashwords™
your ebook. your way.

ftfy 2.0 (fixes text for you)

… テキストデータ内のUnicord関連の問題を修正するライブラリ

SpaCy 3.0

… PythonとCythonベースのオープンソーステキスト解析用トークナイザー

データ整形用ライブラリ

"キャラクターの心情変化"
"ストーリーの状況変化"

本は**貴重な情報源**であり,
豊富な説明力を学習可能

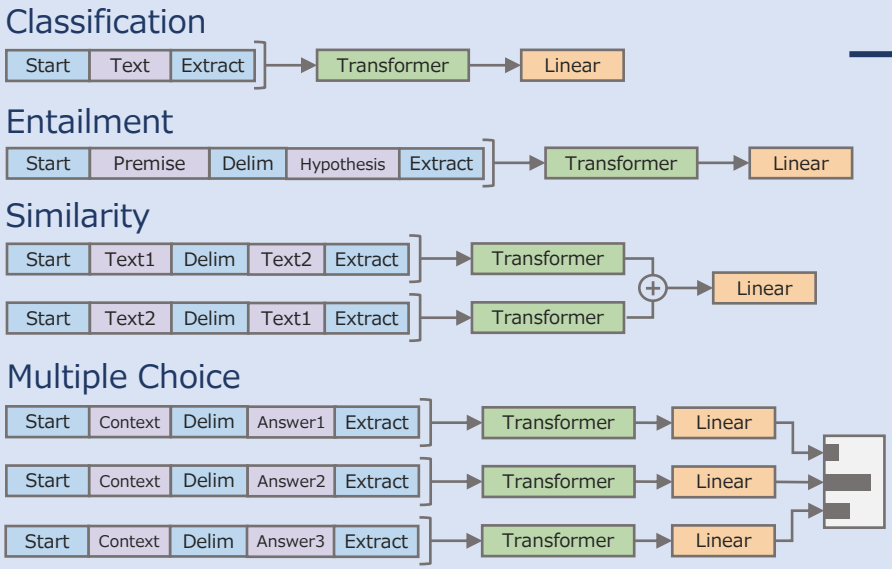
RADFORD, Alec, et al. Improving language understanding by generative pre-training. 2018.

GPTとは

GPT (Generative Pretrained Transformer)

○ Transformerベースの学習済み大規模言語モデル

— 教師なし学習と教師あり学習を組み合わせた学習手法(半教師あり学習)



転移学習ではなく，ファインチューニングによってトラバーサルな学習を実現

教師あり学習フェーズ

ファインチューニング

対象データセット$C = \{x^1, \ldots, x^m\}$を想定

$$P(y|x^1, \ldots, x^m) = softmax(h_l^m W_y)$$

事前学習済みモデルによって出力を得る

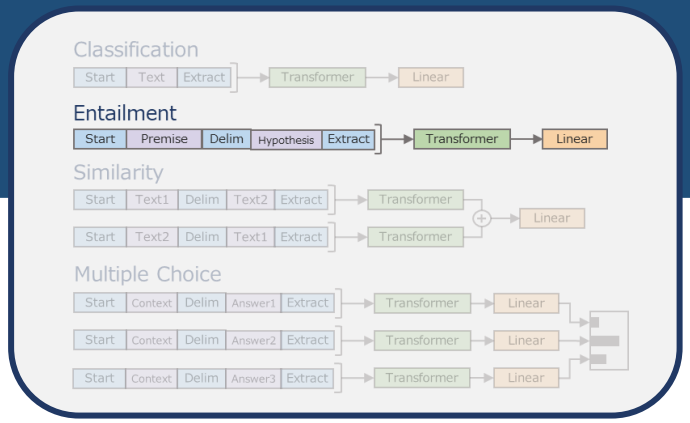
$$L_2(C) = \sum_{(x,y)} \log P(y|x^1, \ldots, x^m)$$

$$L_3(C) = L_2(C) + \lambda * L_1(C)$$

教師なしフェーズと同等の目的関数$L_2(C)$

学習済みのパラメータも含めて学習する

GPTとは



GPT-1の性能評価

○ 自然言語推論(NLI)タスクに対するモデル性能比較

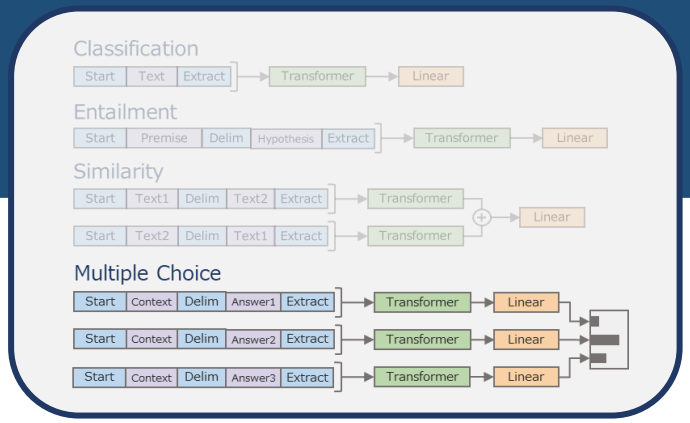
— 5種類のデータセットで評価

- **MNLI**(Multi-Genre Natural Language Interface Matched / MultiNLI Mismatched) … フィクション, 政府の報告書
- **SNLI**(Standard Natural Language Interface) … 画像のキャプションから作成
- **SciTail** … 多肢選択式の科学試験とWeb文章から作成
- **QNLI**(Question NLI) … Wikipediaの一連の記事から作成
- **RTE**(Recognizing Textual Entailment) … ニュース記事から作成

Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM+ELMo(5x)	-	-	89.3	-	-	-
CAFE(5x)	80.2	79.0	89.3	-	-	-
Stohastic Answer Network(3x)	80.6	80.1	-	-	-	-
CAFE	78.7	77.9	88.5	83.3	-	-
GenSen	71.4	71.3	-	-	82.3	59.2
Multi-task BiLSTM+Attn	72.2	72.1	-	-	82.1	**61.7**
Finetuned Transformer LM	**82.1**	**81.4**	**89.9**	**88.3**	**88.1**	56.0

4つのデータセットに対して, 複数文に対する合理的推論や, 言語的曖昧性の処理能力が示された

GPTとは



GPT-1の性能評価

○ 質問応答(Question Answering)タスクに対するモデル性能比較

— **2種類のデータセットで評価**

・**Story Cloze Test** … 複数文から成るストーリーに対して, 2つの選択肢の内, 正しい結末を当てるテスト

例)

Karen was assigned a roommate her first year of college.
Her roommate asked her to go to a nearby city for a concert.
Karen agreed happily. The show was absolutely exhilarating.

▶

A. Karen became good friends with her roommate.

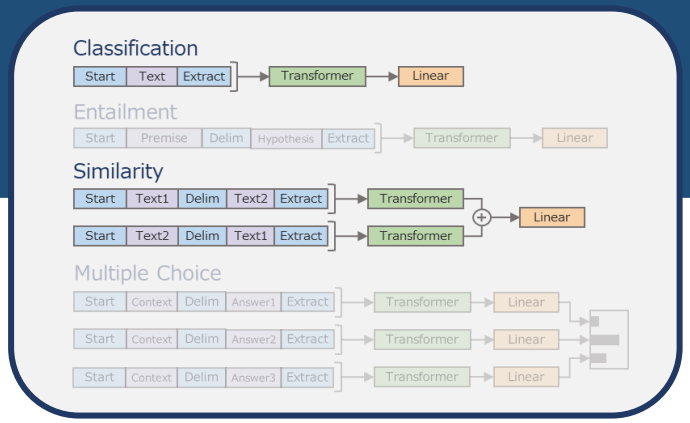
B. Karen hated her roommate.

・**RACE**(ReAding Comprehension Dataset From Examinations) … 中学, 高校の試験問題から作成

Method	Story Cloze	RACE-m	RACE-h	RACE
Val-LS-skip	76.5	-	-	-
Hidden Coherence Model	77.6	-	-	-
Dynamic Fusion Net	-	55.6	49.4	51.2
BiAttention MRU	-	60.2	50.3	53.3
Finetuned Transformer LM	**86.5**	**62.9**	**57.4**	**59.0**

▶ **全てのデータセットに対して, 長文の文脈を効果的に処理可能であることが示された**

GPTとは



GPT-1の性能評価

○ テキスト分類・意味的類似性タスクに対するモデル性能比較

— [テキスト分類]2種類, [意味的類似性]3種類のデータセット + GLUEテストで評価

- **CoLA**(The Corpus of Linguistic Acceptability) … 全23種類の言語学出版物から作成
- **SST2**(The Stanford Sentiment Treebank) … ネガポジ判定用(2値分類)
- **MRPC**(Microsoft Research Paraphrase Corpus) … Web上のニュース記事から作成
- **STS-B**(Semantic Textual Similarity Benchmark) … 計算意味解析システム評価会「SemEval」によって作成
- **QQP**(Quora Question Pairs) … Q&Aサイト「Quora」の質問から作成

Method	Classification		Semantic Similarity			GLUE
	CoLA	SST2	MRPC	STS-B	QQP	
Sparse byte mLSTM	-	**93.2**	-	-	-	-
TF-KLD	-	-	**86.0**	-	-	-
ECNU(mixed ensemble)	-	-	-	81.0	-	-
Single-task BiLSTM+ELMo+Attn	35.0	90.2	80.2	55.5	66.1	64.8
Multi-task BiLSTM+ELMo+Attn	18.9	91.6	83.5	72.8	63.3	68.9
Finetuned Transformer LM	**45.4**	91.3	82.3	**82.0**	**70.3**	**72.8**

3つのデータセットとGLUEベンチマークテストに対して, 大幅な性能向上が示された

GPT-2への進化

GPT-1 → GPT-2で何が変わったのか

○ 多様なタスクに対応可能な汎用的な言語モデルを構築

― 従来のマルチタスクモデルの構築は, “特定のタスクに対して教師ありデータを用いたアプローチ”が主流

$$P(x) = \prod_{i=1}^{n} p(s_n|s_1, \dots, s_{n-1})$$

$$\rightarrow P(s_{n-k}, \dots, s_n|s_1, \dots, s_{n-k-1})$$

$$\rightarrow P(output|input)$$

複数のタスクを単一モデルで解く

$$P(output|input, \boldsymbol{task})$$

― モデルの構造はあまり変化していない (規模は増加)

― 学習データとパラメータ数が大幅に増加

… 学習データ：4.5GB → 40GB (WebText)

… パラメータ数：1億1700万 → 15億4200万

GPT-2への進化

GPT-2のZero-Shot学習用データについて

○ 事前学習用データセット

— **Webのクローリングデータ (WebText)**

… 掲示板型ソーシャルニュースサイト「Reddit」から作成

… 3 karma(高評価のようなモノ)以上獲得している投稿のみに限定

… Wikipediaは, 評価用データセットと重複しているため避ける (リーケージ回避)

約800万文書分
(約40GB)

Text
“I’m not the cleverest man in the world, but like they say in French: Je ne suis pas un imbecile[I’m not a fool].
“I hate the work ‘perfume, ’” Burr says. ‘It’s somewhat better in French: ‘parfum.’
“Brevet Sans Garantie Du Gouvernement”, translated to English: “Patented without government warranty”.

reddit

○ byte-level BPE(Byte Pair Encording)

— 文字列をByte文字列に変換した後BPE圧縮を適用し, 低頻度 / 未知語に対して効率的に対応

例)

ABBBBCCBBCC → AZZCCZCC → AZZYZY → AZXX

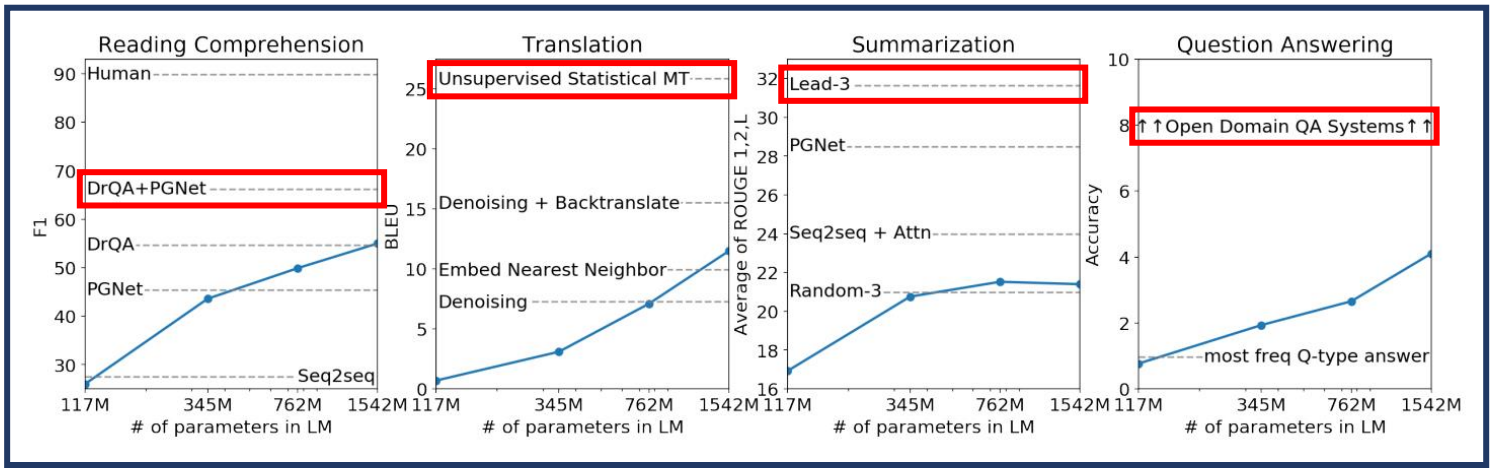
GPT-2への進化

GPT-2の性能評価

○ 読解, 翻訳, 要約, 質疑応答タスクに関するモデル性能比較

― 各タスクの評価用データセット

- [読解] **CoQA**(Conversation Question Answering) … 7種類の会話テキストデータ
- [翻訳] **WMT2014 English-German / German-English** … 2014年度統計的機械翻訳ワークショップで使用
- [要約] **CNN Daily Mail Dataset** … ニュース記事や新聞記事から本文と要約のペアを作成
- [質疑応答] **SQuAD**(The Stanford Question Answering Dataset) … Wikipedia記事から作成



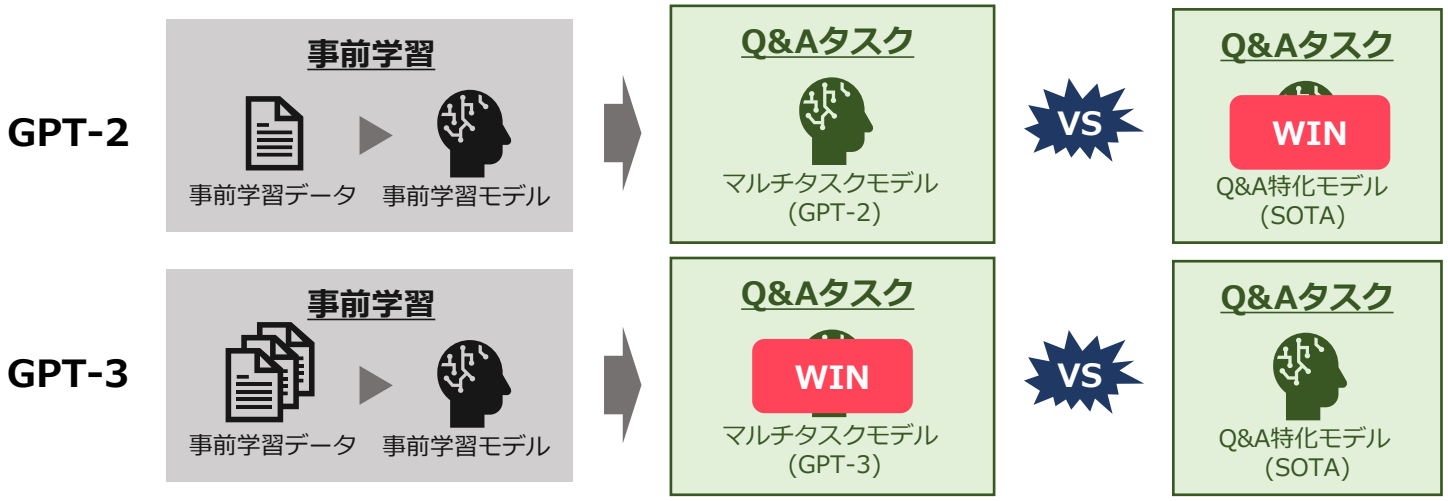
単一Zero-Shotモデルの性能としては高いが, 特定タスク専門モデルにはまだまだ劣る、、

GPT-3への進化

GPT-2 → GPT-3で何が変わったのか

○ GPT-2より大容量データで, より大規模モデルを学習

― タスク特化モデルより性能が良いマルチタスクモデルの構築が目的



― 学習データとパラメータ数が大幅に増加

… 学習データ：40GB (WebText) → 570GB (Common Crawl, 書籍, WebText等)
… パラメータ数：15億4200万 → 1750億

― 当然, GPT-3をファインチューニングすれば, 非常に高い精度が期待できる (本来の趣旨ではない)

GPT-3への進化

GPT-3の性能評価 (単語予測)

○ PTB, LAMBADAデータセットを用いて，言語モデル精度評価

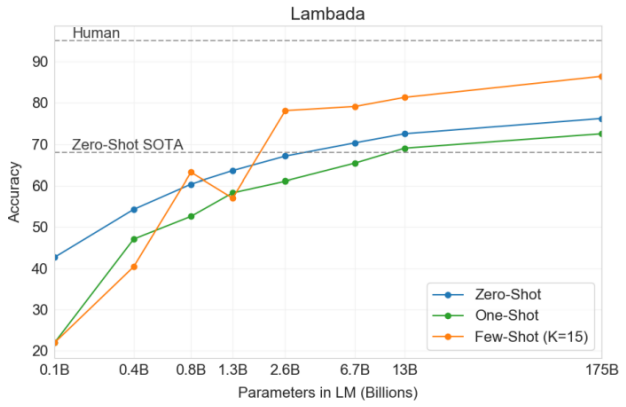
— 評価指標：PPL … 言語モデルの良し悪しを評価する指標の一つ

$$ppl = \exp(-\log(\text{"True Word Prediction Probability"})) = \frac{1}{\text{"True Word Prediction Probability"}}$$

正解単語の選択肢数が
何択まで絞れているのかを
表す指標とも解釈可能

— 任意の文の最後の単語を予測するタスク

Method	PTB (PPL)	LAMBADA (PPL)	LAMBADA (ACC)
SOTA (GPT-2)	35.8	8.63	68.0
GPT-3 (Zero-Shot)	**20.5**	**3.00**	**76.2**
GPT-3 (One-Shot)	-	**3.35**	**72.5**
GPT-3 (Few-Shot)	-	**1.92**	**86.4**



縦軸：Accuracy, 横軸：パラメータ数

単語予測タスクにおけるSOTAを超える性能を達成
このタスクは，Zero-Shotと相性が良い？(データ量でゴリ押せる)

GPT-3への進化

GPT-3の性能評価 (質問応答・翻訳)

〇 質問応答タスク

ー

〇 翻訳タスク

ー