注意モデルの変遷と展開

浅川伸一 (東京女子大学) asakawa@ieee.org

14/Jun/2020

エピグラフ

蕉門に千歳不易の句、一時流行の句と云ふ有り。
是を二つに分て教へ給へる、其元は一つ也。
不易を知らざれば基たちがたく、流行を知らざれば風新たならず

— 去来抄

found-his-socks-what-a-great-monday}

自己紹介

師匠エルマンと USCDにて

浅川伸一:博士(文学) 東京女子大学情報処理センター勤務。早稲田大学在学時はピアジェの発生論的認識論に心酔する。卒 業後エルマンネットの考案者ジェフ・エルマンに師事, 薫陶を受ける。以来人間の高次認知機能をシミュレートすることを 通して知的であるとはどういうことかを考えていると思っていた。 著書に 「AI白書 2019, 2018」(2019年, アスキー出版, 共著), 「深層学習教科書 ディープラーニング G検定（ジェネラリスト） 公式テキスト」(2018年, 翔泳社, 共著), 「Python で体験する深層学習」(コロナ社, 2016), 「ディープラーニング, ビッグデータ, 機械学習あるいはその心理学」(新曜社, 2015), 「ニューラルネットワークの数理的基礎」「脳損傷とニューラル ネットワークモデル, 神経心理学への適用例」いずれも守一雄他編「コネクショニストモデルと心理学」(2001)北大路書房 など

アウトライン

1. どこにでも現れる注意
2. BERT 概説
3. 流行の句あり
4. 不易の句あり
5. まとめ

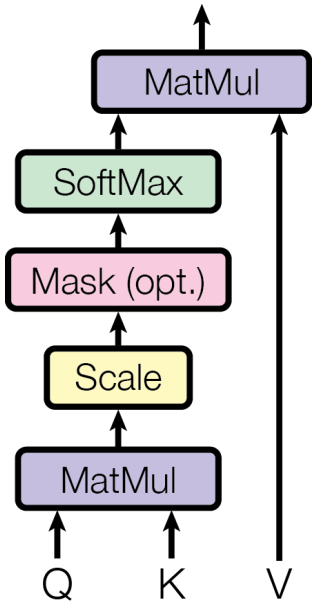
第 1 部

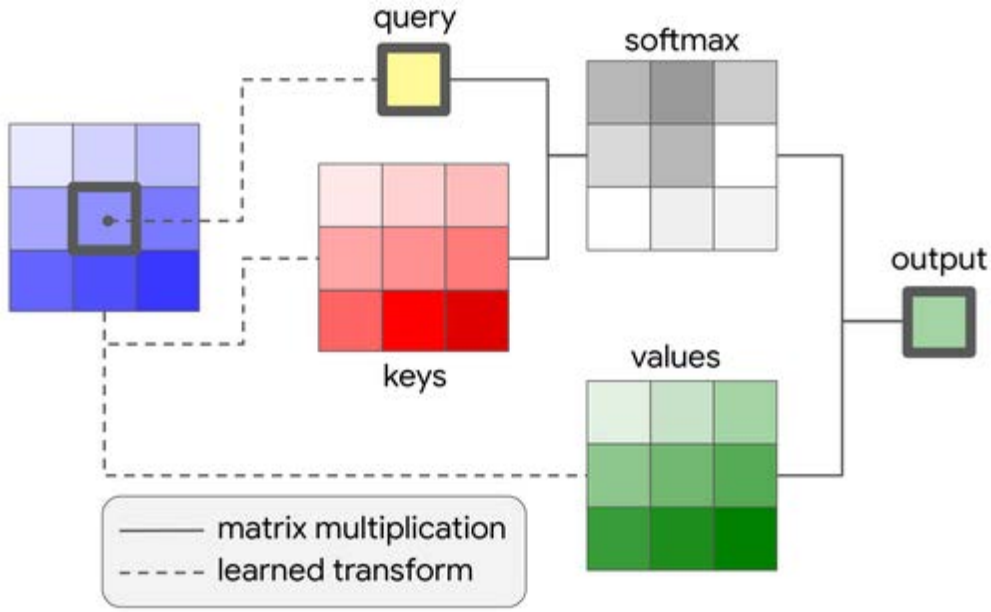
どこにでも現れる注意

多頭=自己注意 Multi-Head Self-Attention: MHSA

- 自然言語処理 NLP **Transformer**(Vaswani et al. 2017); **BERT**(Devlin et al. 2018); **RoBERTa**(Y. Liu et al. 2019); **distilBERT** (Sanh et al. 2020); and more…
- 画像処理 (Ramachandran et al. 2019); **A2-Net** (Chen et al. 2018); **U-GAT-IT** (Kim et al. 2019)
- 強化学習，メタ学習 **SNAIL** (Mishra et al. 2018)
- 敵対生成ネットワーク **SAGAN** (Zhang et al. 2019)

多頭=自己注意 Multi-Head Self-Attention





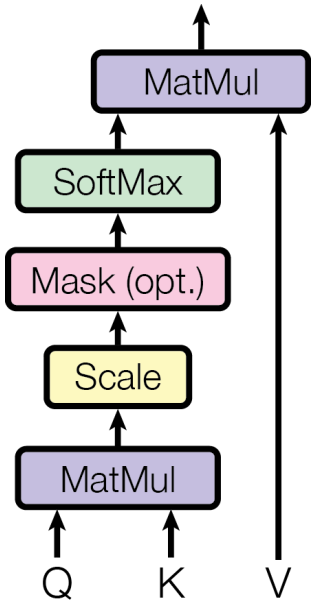
Left: (Vaswani et al. 2017), Right: (Ramachandran et al. 2019)

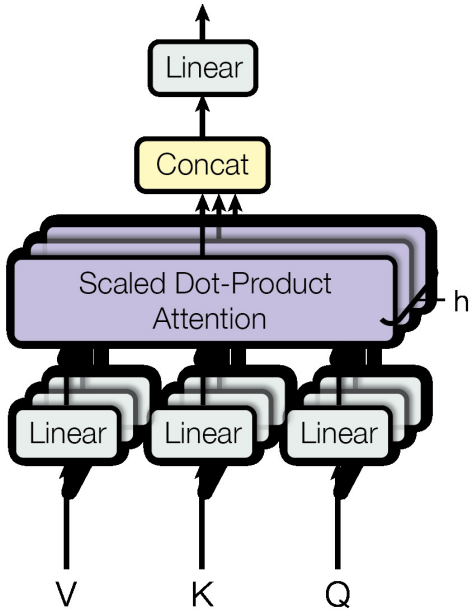
$$\text{自己注意}\left(\mathbf{X}_{t,:}\right) = \text{ソフトマックス}\left(\mathbf{A}_{t,:}\right)\mathbf{X}\mathbf{W}_{\text{バリ}}^{(1}$$

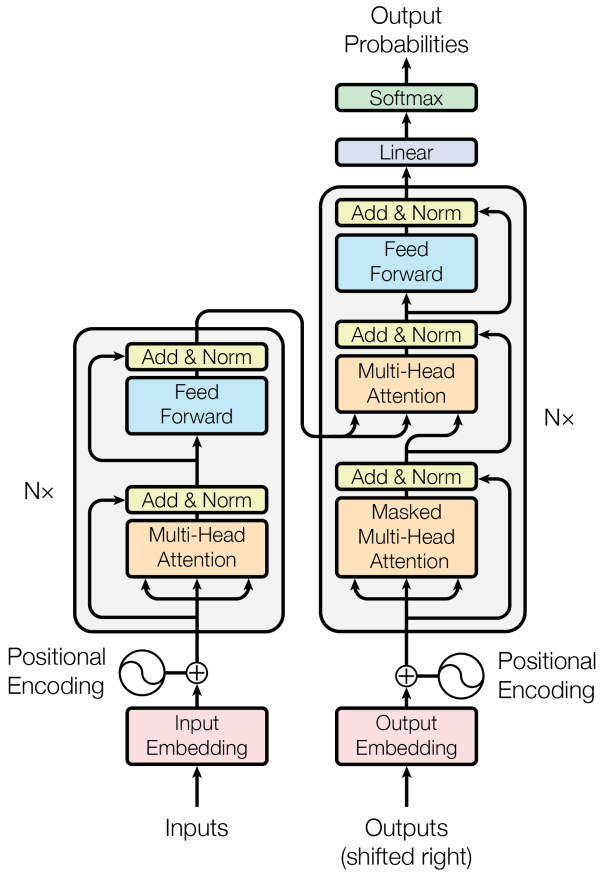
$$\mathbf{A} = \mathbf{X}\mathbf{W}\mathbf{W}_{\text{クエキクk}}^{\top}\mathbf{X}^{\top} \qquad (2$$

$$\mathbf{A} := (\mathbf{X}+\mathbf{P})\,\mathbf{W}\mathbf{W}_{\text{クエキクk}}^{\top}(\mathbf{X}+\mathbf{P})^{\top}, \qquad \mathbf{P}\text{ は 位置符号化器PE} \qquad (3$$

Multi-head self-attention: MHSA (2)







謝辞

本日，このような機会を与えていただきましたエクサウィザーズ，藤井 亮宏 様，遠藤 太一郎 様 に感謝申し上げます。

ABBA - The Winner Takes It All (1980) HD...

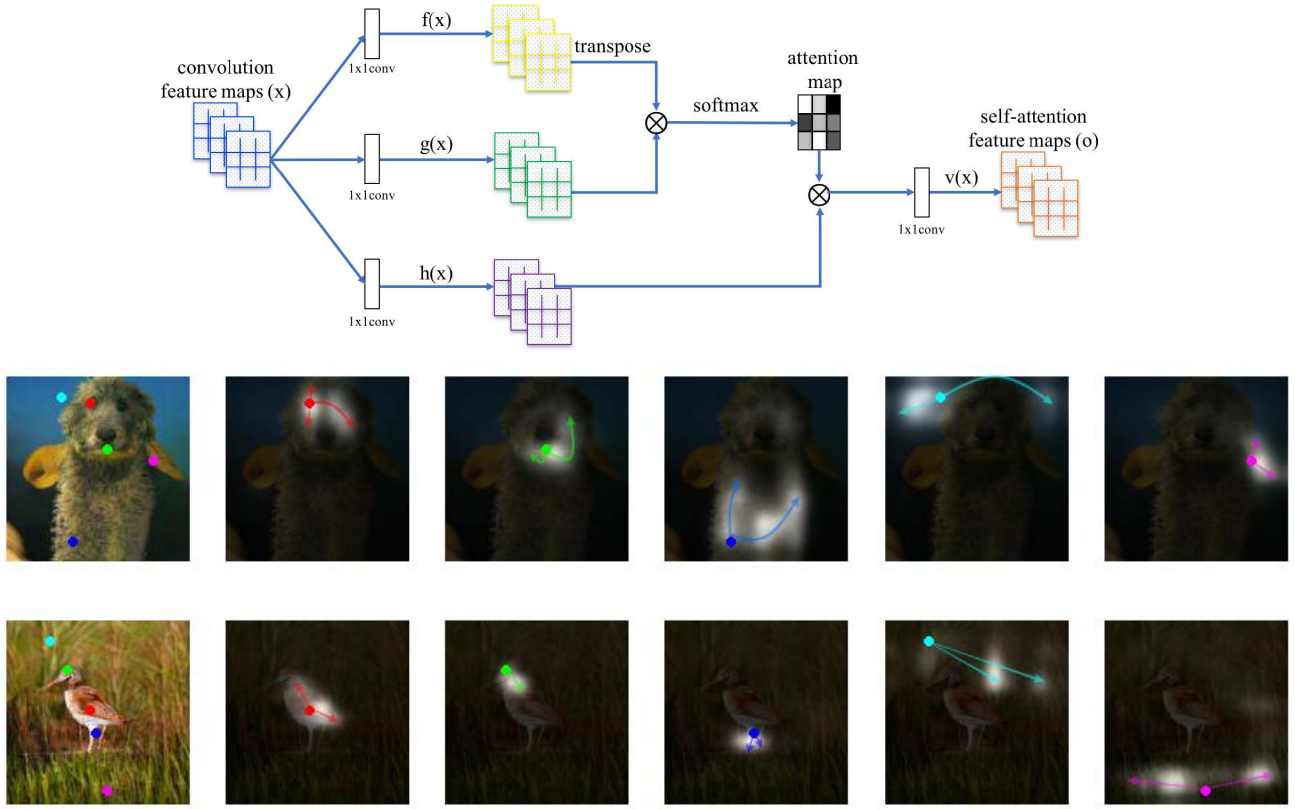
Love Is All You Need - Beatles

トランスフォーマーOP

credit: {‘Beatles’:https://youtu.be/dsxtImDVMig, ABBA:https://youtu.be/92cwKCU8Z5c, ‘トランスフォーマー’: https://youtu.be/cwpXeH90qfE), ‘ELMO’:

https://giftsandwish.com/christmas-gifts-for-kids/playskool-friends-sesame-street-tickle-me-elmo/, ‘BERT’: https://sesamestreet.tumblr.com/post/5772176064/bert-

Multi-head self-attention: MHSA (3) SAGAN (Self-Attention GAN)

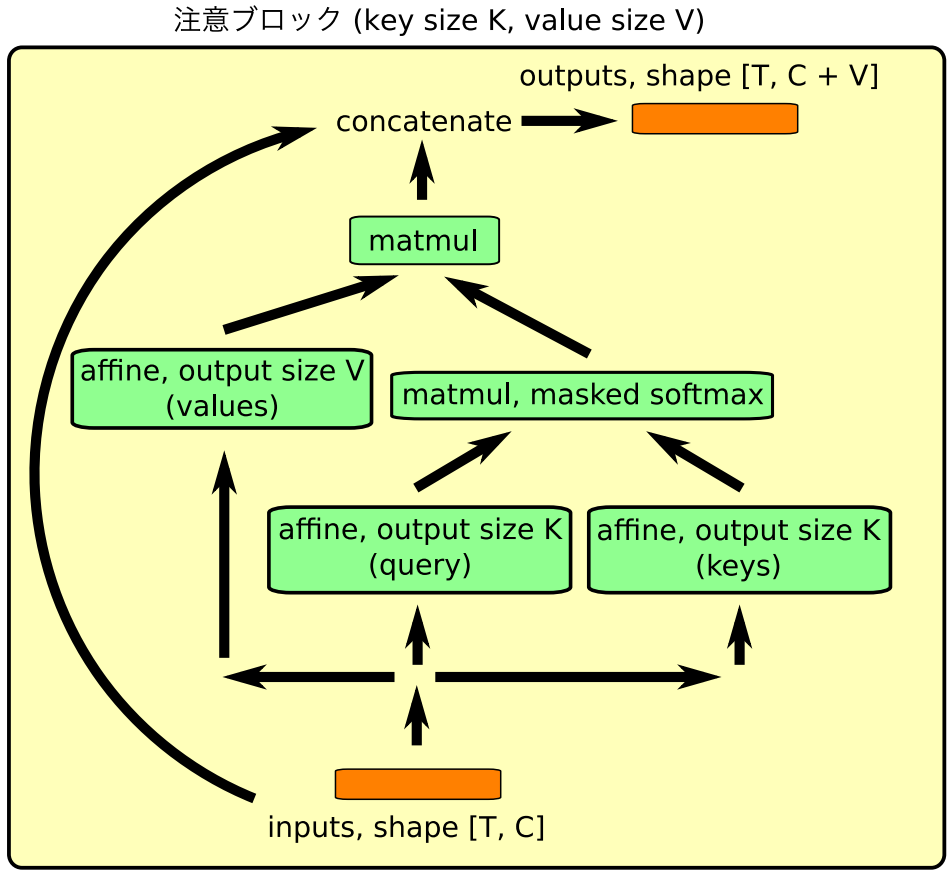


From (Zhang et al. 2019) Fig. 1, and 3. 画像生成において，近傍画素から情報だけでなく，関連する遠距離の特徴を利用して生成することにより一貫性のある対象やシナリオを生成可能。各行の左の元画像上のカラー点は 5 つ の 代表的なクエリの場所を示す。右側の 5 画像は 各クエリ位置における注意地図。最も注目されている領域が，色分けされた矢印で示されている。

Multi-head self-attention: MHSA (4) Non-Local Net

時空の非局所ネットワークの概念図。特徴地図はテンソルとして示されている。 例えば 1024 チャンネルの場合は $T \times H \times W$ 行列積を，$\oplus$ は要素和を示す。 ソフトマックス演算は各行に対して実行される。 青いボックスは $1 \times 1 \times 1$ の畳み込みを表す。 512 チャンネルのボトルネックを持つ埋め込みガウシアン版が示されている。 バニラガウス版は θ と ϕ を除去することで ドット積版は $1/N$ スケーリングでソフトマックスを置き換えることで行うことができる。 From (X. Wang et al. 2018)

Multi-head self-attention: MHSA (4) SNAIL



From (Mishra et al. 2018) Fig. 2

トランスフォーマーはリカレント構造や畳み込み構造を持たず埋め込みベクトルに位置符号化器を加えることで系列情報を処理する。 しかし、逐次的な順序情報が貧弱であるとの批判がある。 とりわけ強化学習のような位置依存性に敏感な課題では問題。 トランスフォーマーモデルにおける 位置問題を解決するため, 自己注意機構 と 時間的な畳み込み temporal convolution を組み合わせたモデ

ルが Simple Neural Attention Meta-Learner (SNAIL)(Mishra et al. 2018)。 SNAIL は，メタ学習，強化学習の両方の課題に優れていることが実証された。

注意用語集 Taxosonomy of attention

- **文脈ベース** 注意 context-base attention: $\mathrm{score}(\mathbf{s}_t, \mathbf{h}_i) = \cos(\mathbf{s}_t, \mathbf{h}_i)$ (Graves, Wayne, and Danihelka 2014)
- **加算的** (連結的) 注意 Additive : $\mathrm{score}(\mathbf{s}_t, \mathbf{h}_i) = \mathbf{v}_a^\top \tanh(\mathbf{W}_a[\mathbf{s}_t; \mathbf{h}_i])$ (Bahdanau, Cho, and Bengio 2015)
  - *(Luong, Pham, and Manning 2015) では 連結 concatenated, (Vaswani et al. 2017) では 加算 additive と表記されている*
- **場所ベース** 注意 Location-Base: $\alpha_{t,i} = \mathrm{softmax}(\mathbf{w}_a \mathbf{s}_t)$ (Luong, Pham, and Manning 2015)
- Note: This simplifies the softmax alignment to only depend on the target position.
- **一般的** 注意 general: $\mathrm{score}(\mathbf{s}_t, \mathbf{h}_i) = \mathbf{s}_t^\top \mathbf{W}_a \mathbf{h}_i$ (Luong, Pham, and Manning 2015)
- $\mathbf{W}_a$学習可能な結合係数行列
- **ドット積** 注意 dot-product: $\mathrm{score}(\mathbf{s}_t, \mathbf{h}_i) = \mathbf{s}_t^\top \mathbf{h}_i$ (Luong, Pham, and Manning 2015)
- **スケール化ドット積** 注意 scaled dot-product(^): $\mathrm{score}(\mathbf{s}_t, \mathbf{h}_i) = \frac{\mathbf{s}_t \mathbf{h}_i}{\sqrt{n}}$ (Vaswani et al. 2017)
  - *スケール化規格化因子 $\frac{1}{\sqrt{n}}$を用いる*

第1部 Multi-head self-attention: MHSA のまとめ

- 自然言語処理，画像処理，強化学習，メタ学習の 4 分野でほぼ同様の MHSA が取り入れられている。
- クエリ，キー，バリュー 各テンソルを学習することが行われている
- 従来手法である 畳み込み や LSTM を MHSA で置き換える動きがある。
- ただし, SAGAN と SNAIL （non-local net） では 入力情報を concatenate して上位層に伝える点が他と異なる

補足 注意が現れるに至った歴史

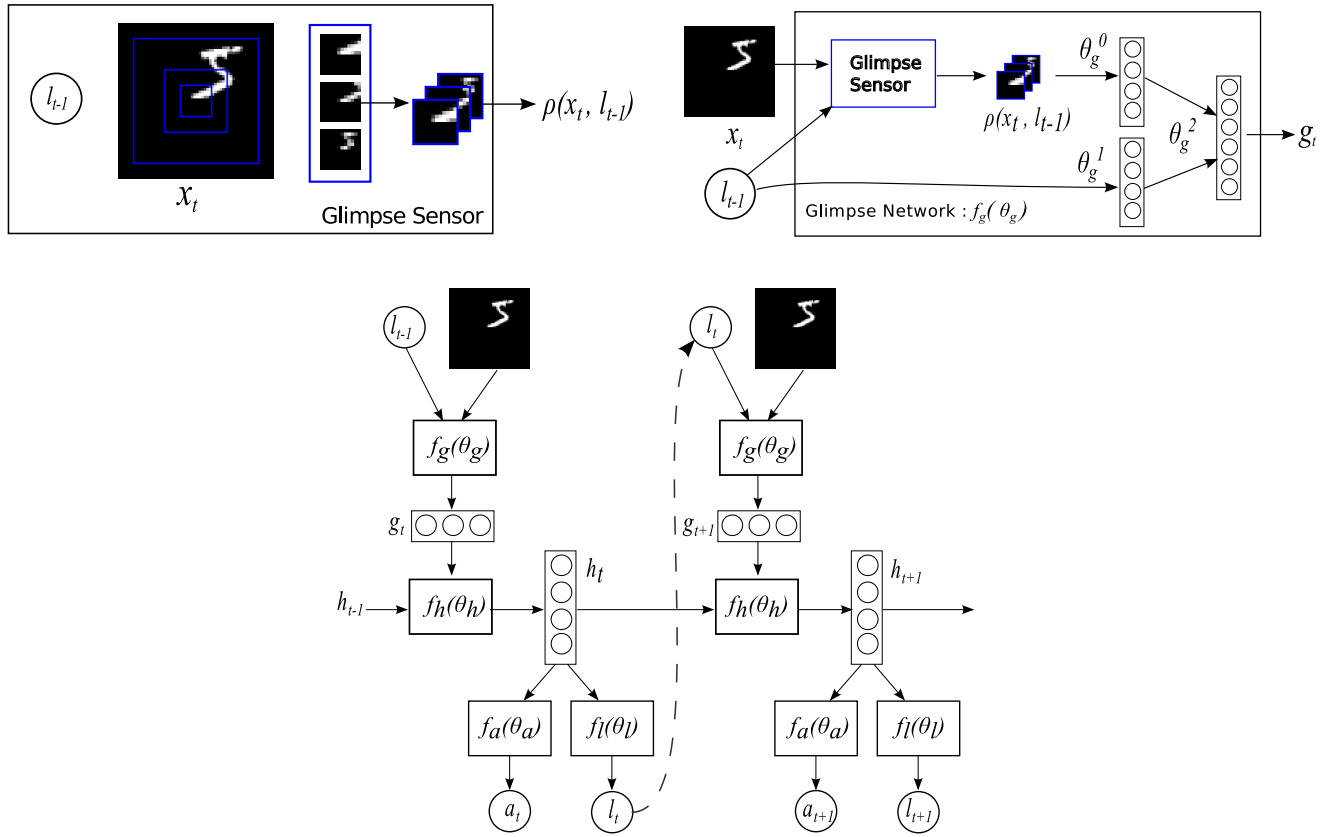
- BOW, TFIDF(Jones 1972), SMT(Manning and Schuütze 1999), N-gram モデル, Dimensionality would increse w.r.t. V^N
- RNN (Elman 1990),(Mikolov et al. 2010)(Mikolov et al. 2011)
- LSTM (Hochreiter and Schmidhuber 1997),(Gers, Schmidhuber, and Cummins 1999),(Greff et al. 2015), **Seq2seq**(Sutskever, Vinyals, and Le 2014), 注意モデル(Bahdanau, Cho, and Bengio 2015),
- Transformer (Vaswani et al. 2017)
- BERT (Devlin et al. 2018)

それぞれ有名なので説明はしません

第 2 部

BERT 概説

Mnih and Graves (2014)



From (Mnih et al. 2014)

Show and Tell (2014)

Figure 2. Attention over time. As the model generates each word, its attention changes to reflect the relevant parts of the image. "soft" (top row) vs "hard" (bottom row) attention. (Note that both models generated the same captions in this example.)

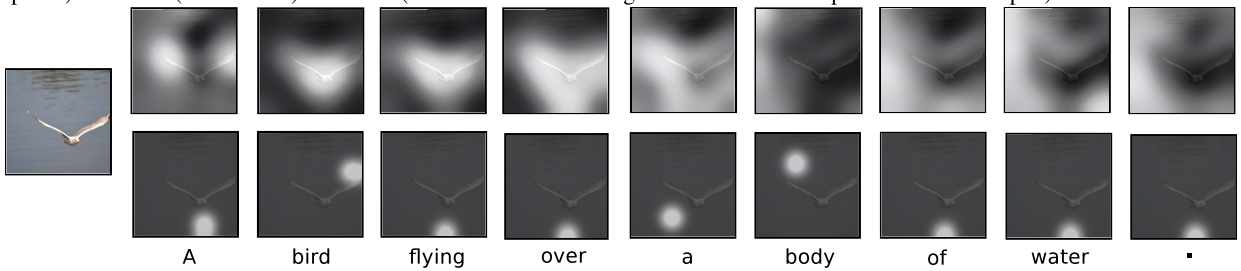
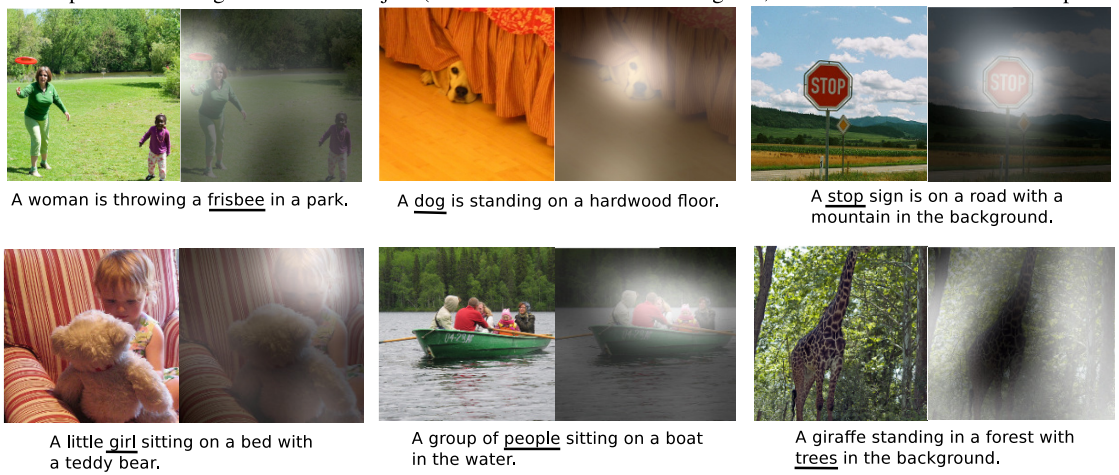
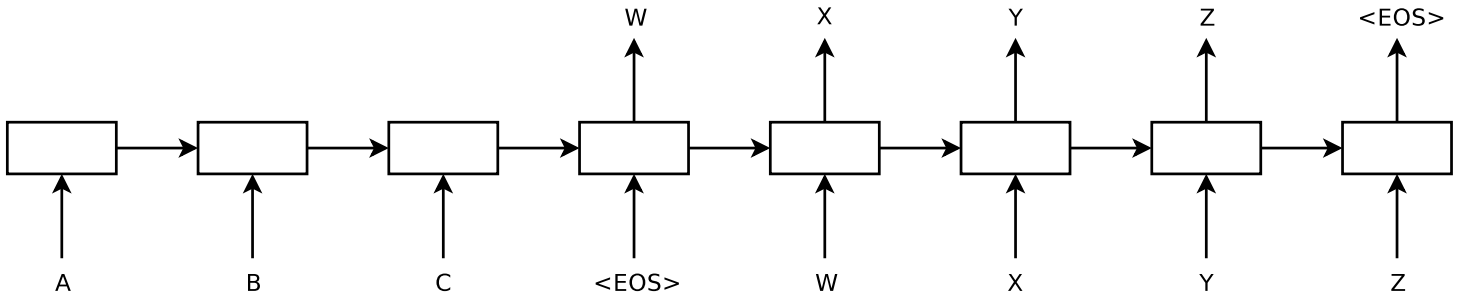


Figure 3. Examples of attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word)



Attention for neural image captioning (Xu et al. 2015)

Seq2seq model

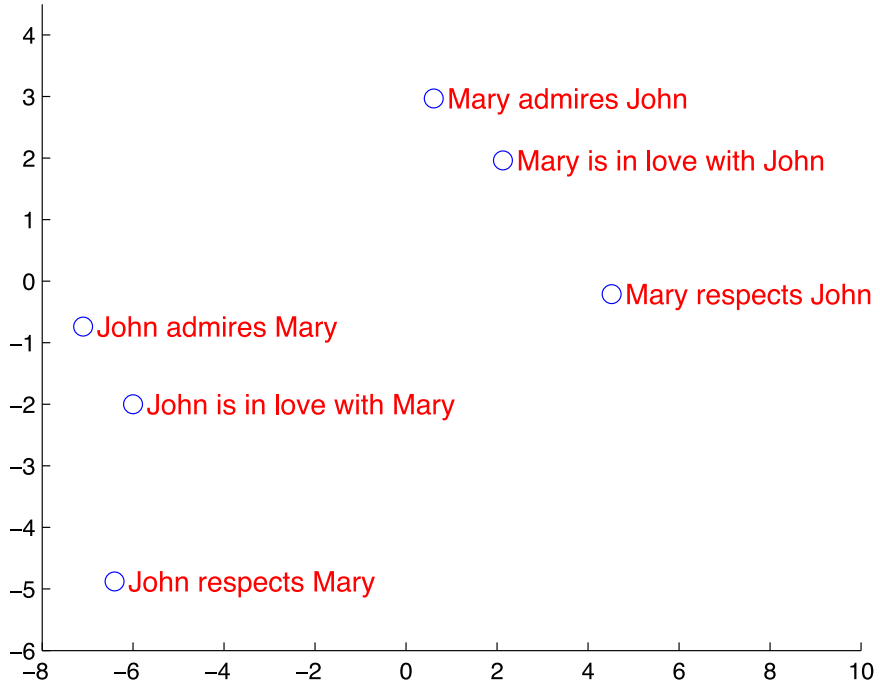


From (Sutskever, Vinyals, and Le 2014) Fig. 1, 翻訳モデル “seq2seq” の概念図

“eos” は文末を表す。中央の “eos” の前がソース言語であり，中央の “eos” の後はターゲット言語の言語モデルである SRN の中間層への入力として用いる。

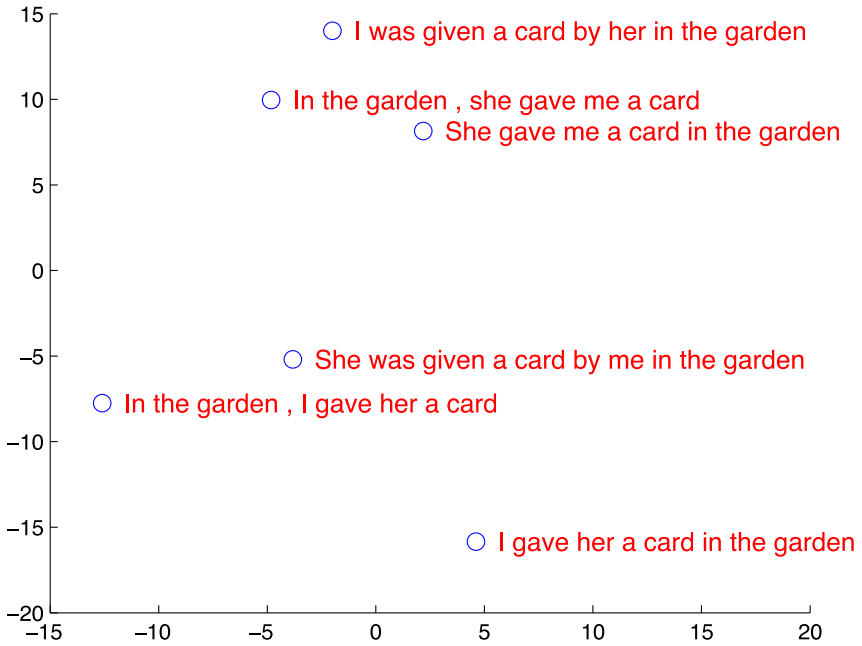
注意すべきは，ソース言語の文終了時の中間層状態のみをターゲット言語の最初の中間層の入力に用いることであり，それ以外の時刻ではソース言語とターゲット言語は関係がない。逆に言えば最終時刻の中間層状態がソース文の情報全てを含んでいるとみなしうる。この点を改善することを目指すことが 2014 年以降盛んに行われてきた。顕著な例が後述する **双方向 RNN**，**LSTM** 採用したり，**注意** 機構を導入することであった。

Seq2seq (2)



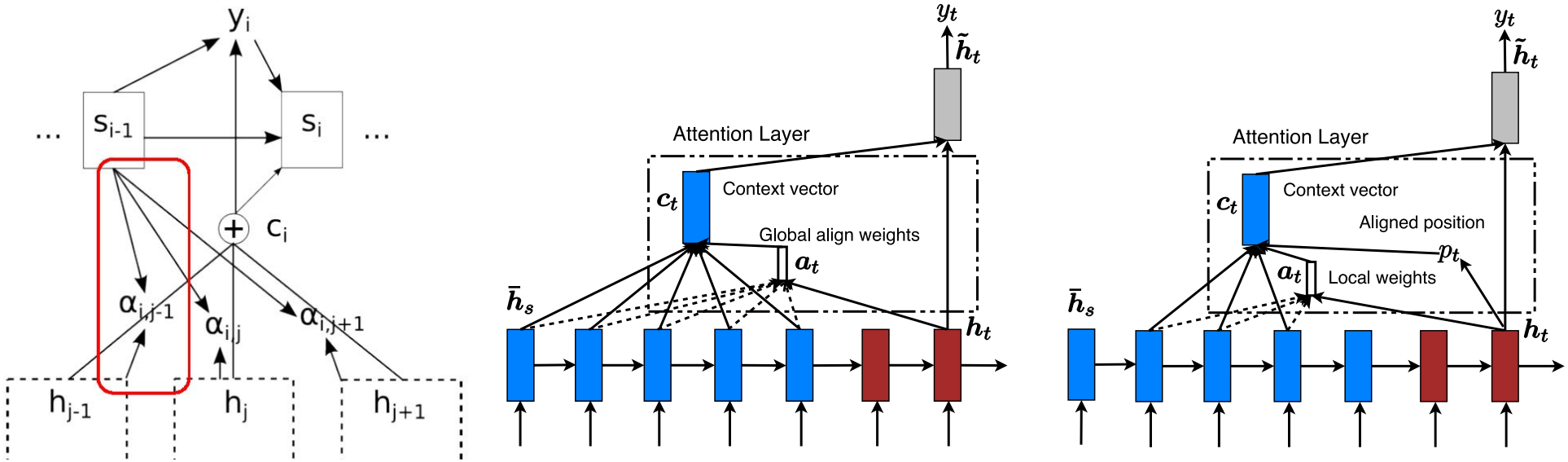
From (Sutskever, Vinyals, and Le 2014) Fig. 2

Seq2seq (3)



From [Sutskever, Vinyals, and Le (2014)} Fig. 2

自然言語系の注意



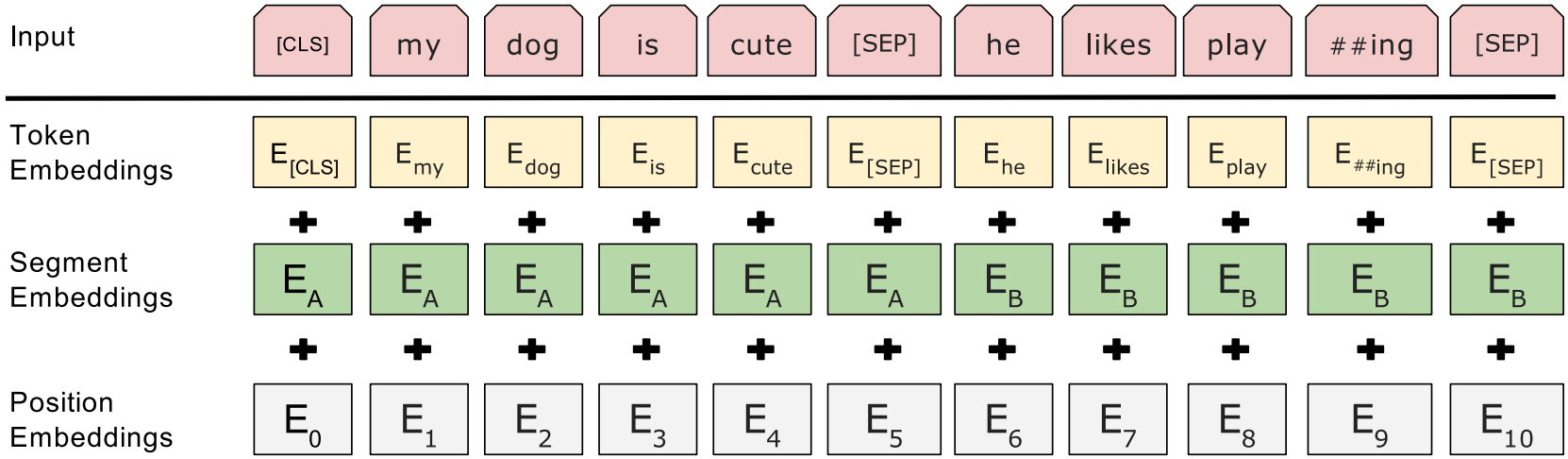
左:[Bahdanau, Cho, and Bengio (2015)}, 中:[Luong, Pham, and Manning (2015)} Fig. 2, 右:[Luong, Pham, and Manning (2015)} Fig. 3

BERT の特徴

BERT の特徴を 3 つにまとめると以下の通り

1. トランスフォーマー Transformer に基づく MHSA を用いた多層ニューラルネットワークモデル
2. 2 つの事前訓練: **マスク化言語モデル** と **次文予測課題**
3. Fine tuning によるマルチタスクで性能向上 GLUE スコアボード, SuperGLUE を参照のこと

BERT の入力表現



埋め込みトークンの総和，位置符号器，分離埋め込みの 3 者 From (Devlin et al. 2018) Fig. 2

BERT の事前訓練: マスク化言語モデル

全入力系列のうち 15% をランダムに [MASK] トークンで置き換える

- 入力はオリジナル系列を [MASK] トークンで置き換えた系列
- ラベル: オリジナル系列の [MASK] 部分にの正しいラベルを予測
- 80%: オリジナル入力系列を [MASK] で置換
- 10%: [MASK] の位置の単語をランダムな無関連語で置き換える
- 10%: オリジナル系列

BERT の事前訓練: 次文予測課題

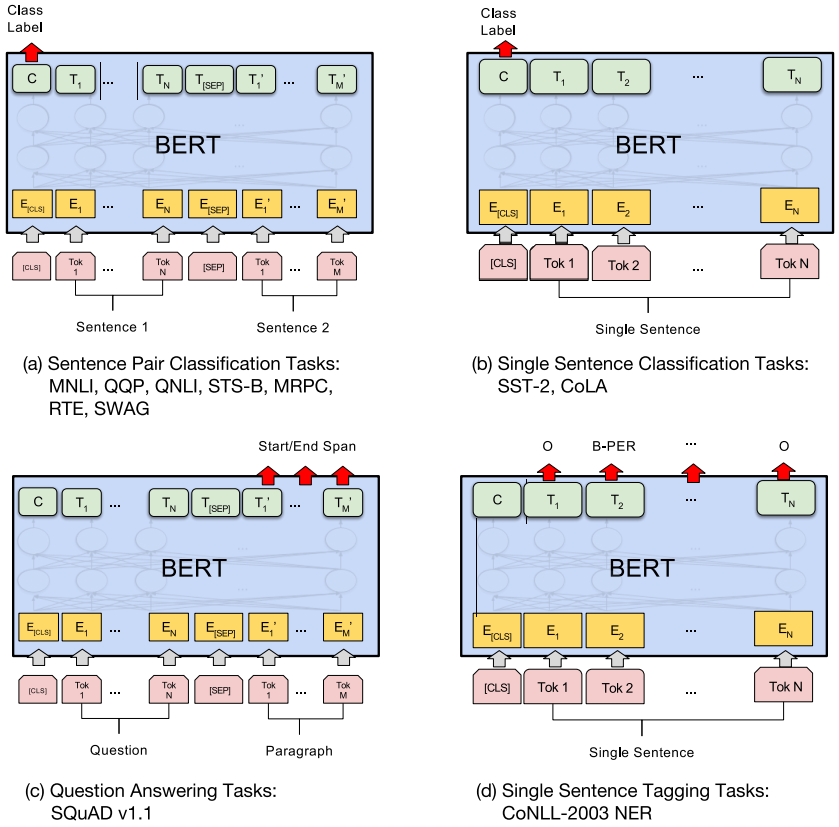
言語モデルの欠点を補完する目的，次の文を予測

[SEP] トークンで区切られた 2 文入力

- 入力: the man went to the store [SEP] he bought a gallon of milk.
- ラベル: IsNext
- 入力: the man went to the store [SEP] penguins are flightless birds.
- ラベル: NotNext

BERT: ファインチューニング

(a), (b) は文レベル課題,　(c),(d)はトークンレベル課題, E: 入力埋め込み表現, T_iトークン iの文脈表象。



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

From (Devlin et al. 2018) Fig.3

GLUE: General Language Understanding Evaluation

- **CoLA**: 入力文が英語として正しいか否かを判定
- **SST-2**: スタンフォード大による映画レビューの極性判断
- **MRPC**: マイクロソフトの言い換えコーパス。2文 が等しいか否かを判定
- **STS-B**: ニュースの見出し文の類似度を5段階で評定
- **QQP**: 2 つの質問文の意味が等価かを判定
- **MNLI**: 2 入力文が意味的に含意，矛盾，中立を判定
- **QNLI**: Q and A
- **RTE**: MNLI に似た2つの入力文の含意を判定
- **WNI**: ウィノグラッド会話チャレンジ

その他

- **SQuAD**: スタンフォード大による Q and A ウィキペディアから抽出した文
- **RACE**: 中学入試，高校入試に相当するテスト多肢選択回答 # BERT モデルの詳細
- データ: Wikipedia (2.5B words) + BookCorpus (800M words)
- バッチサイズ: 131,072 words (1024 sequences * 128 length or 256 sequences * 512 length)

- 訓練時間: 1M steps (~40 epochs)
- 最適化アルゴリズム: AdamW, 1e-4 learning rate, linear decay
- BERT-Base: 12 層, 各層 768 ニューロン, 12 多頭注意
- BERT-Large: 24 層, 各層 1024 ニューロン, 16 多頭注意
- 4x4 / 8x8 TPU で 4 日間

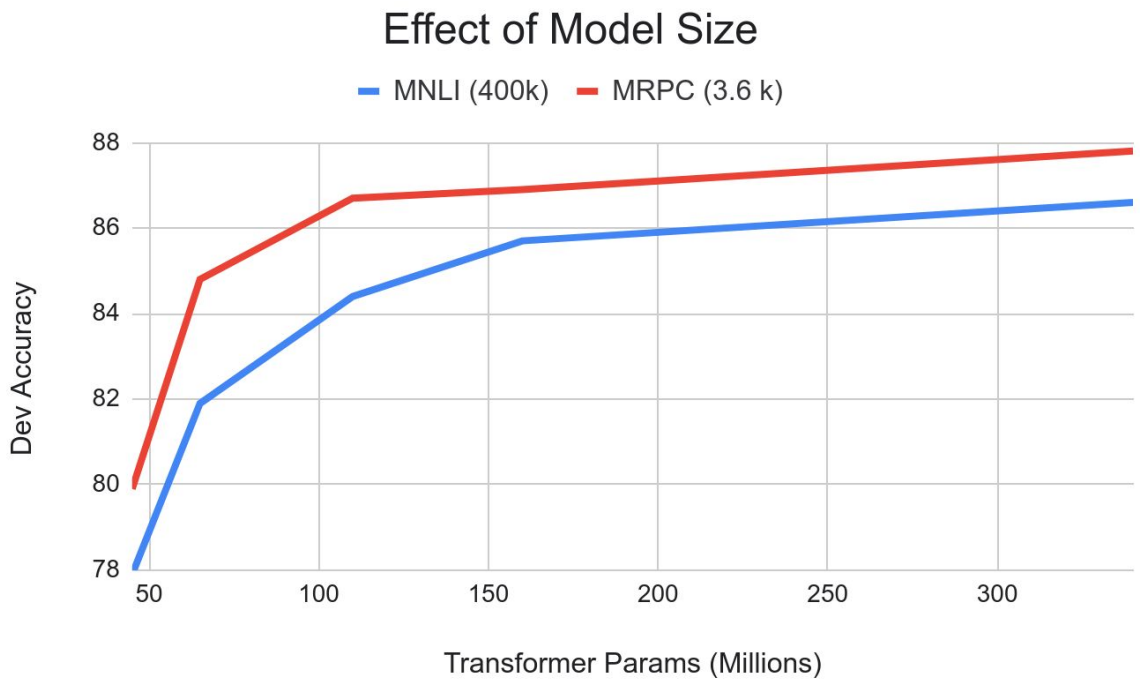
BERT: ファインチューニング手続きによる性能比較

Masking Rates			Dev Set Results		
MASK	SAME	RND	MNLI	NER	
			Fine-tune	Fine-tune	Feature-based
80%	10%	10%	84.2	95.4	94.9
100%	0%	0%	84.3	94.9	94.0
80%	0%	20%	84.1	95.2	94.6
80%	20%	0%	84.4	95.2	94.7
0%	20%	80%	83.7	94.8	94.6
0%	0%	100%	83.6	94.9	94.6

マスク化言語モデルのマスク化割合の違いによる性能比較

マスク化言語モデルのマスク化割合は マスクトークン:ランダム置換:オリジナル=80:10:10 だけでなく， 他の割合で訓練した場合の 2 種類下流課題， MNLI と NER で変化するかを下図 に示した。80:10:10 の性能が最も高いが大きな違いがあるわけではないようである。

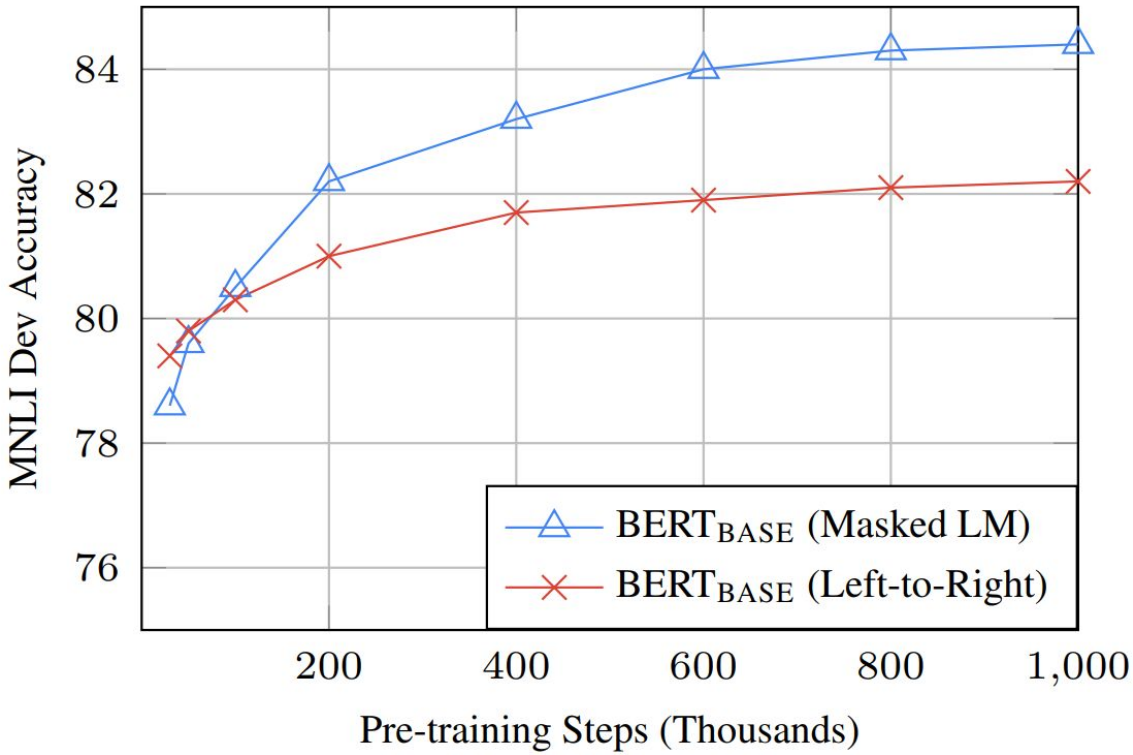
BERT: モデルサイズ比較



モデルのパラメータ数による性能比較

パラメータ数を増加させて大きなモデルにすれば精度向上が期待できる。 下図では，横軸にパラメータ数で MNLI は青と MRPC は赤 で描かれている。 パラメータ数増加に伴い精度向上が認められる。 図に描かれた範囲では精度が天井に達している訳ではない。パラメータ数が増加すれば精度は向上していると認められる。

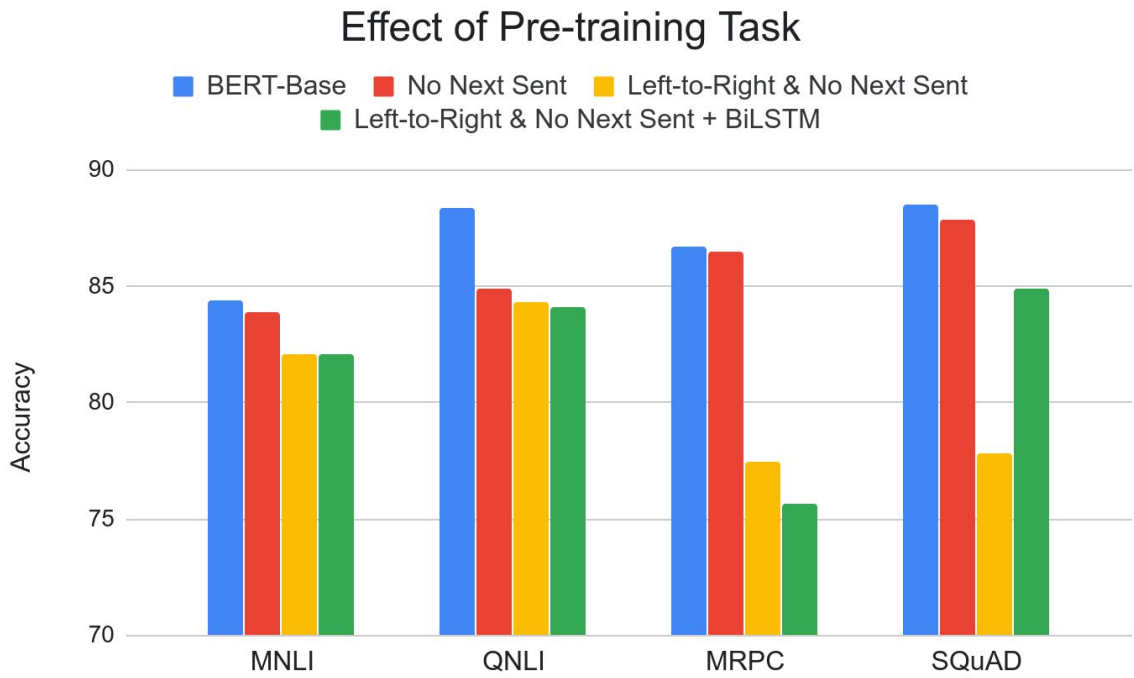
BERT: モデル単方向，双方向モデル比較



言語モデルの相違による性能比較

言語モデルをマスク化言語モデルか次単語予測の従来型の言語モデルによるかの相違による性能比較を 下図 に示した。 横軸には訓練ステップである。訓練が進むことでマスク化言語モデルとの差は２パーセントではあるが認められるようである。

BERT: 事前訓練比較



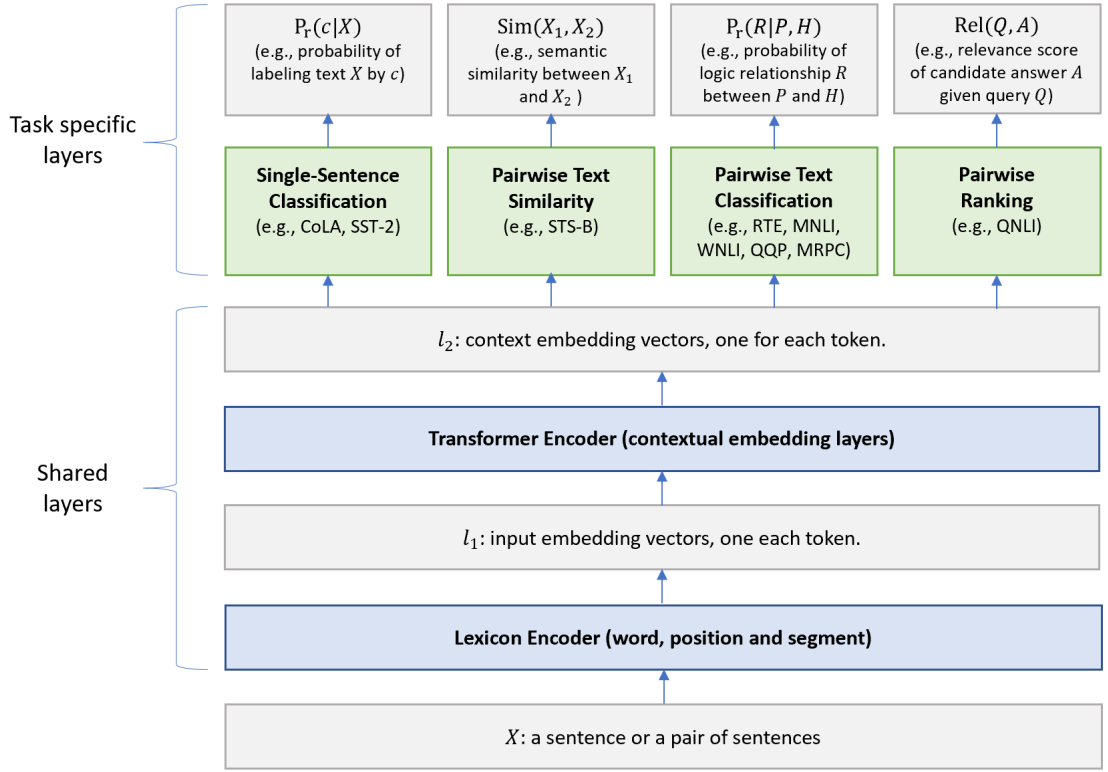
事前訓練の効果比較

図には事前訓練の比較を示しされている。 全ての事前訓練を用いた場合が青, 次文訓練を除いた場合が赤, 従来型言語モデルで次文予測課題をした場合を黄, 従来型言語モデルで次文予測課題なしを緑で描かれている。4 種類の下流課題は MNLI, QNLI, MRPC, SQuAD である。 下流のファインチューニング課題ごとに精度が分かれるようである。

各モデルの特徴

- RoBERTa: BERT の訓練コーパスを巨大 (173GB) にし，ミニバッチサイズを大きした
- XLNet: 順列言語モデル。2 ストリーム注意
- MT-DNN: BERT ベース の転移学習に重きをおいたモデル
- GPT-2: BERT に基づく。人間超えして 2019 年 2 月時点で炎上騒ぎ
- BERT: Transformerに基づく言語モデル。**マスク化言語モデル** と **次文予測** に基づく 事前訓練，各下流課題をファインチューニング。事前訓練されたモデルは一般公開済。
- DistillBERT: BERT の蒸留版
- ELMo: 双方向 RNN による文埋め込み表現
- Transformer: 自己注意に基づく言語モデル。多頭注意，位置符号器.

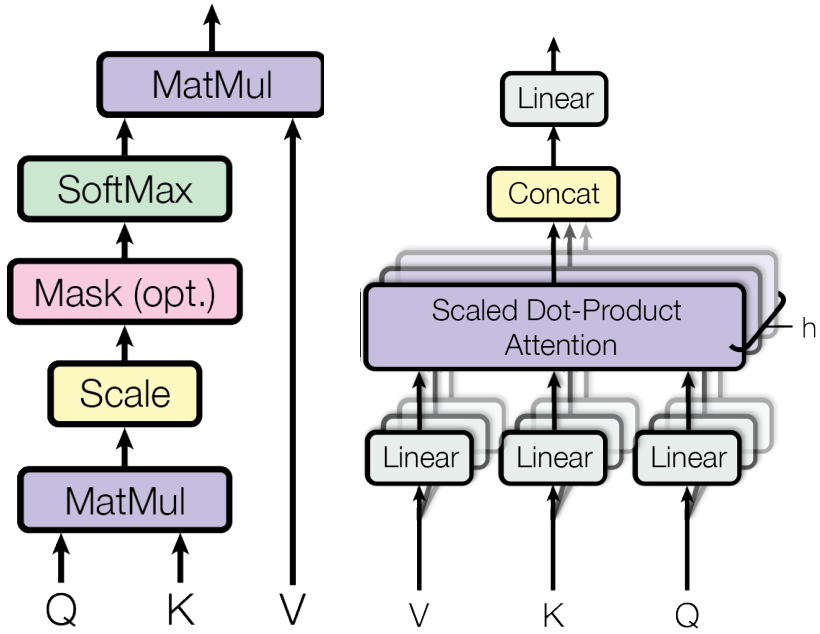
事前訓練とマルチ課題学習



From (X. Liu et al. 2019) Fig. 1

Transformer: Attention is all you need

$$attention(Q, K, V) = dropout\left(softmax\left(\frac{QK^\top}{\sqrt{d}}\right)\right)V \quad (4$$



From (Vaswani et al. 2017) Fig. 2

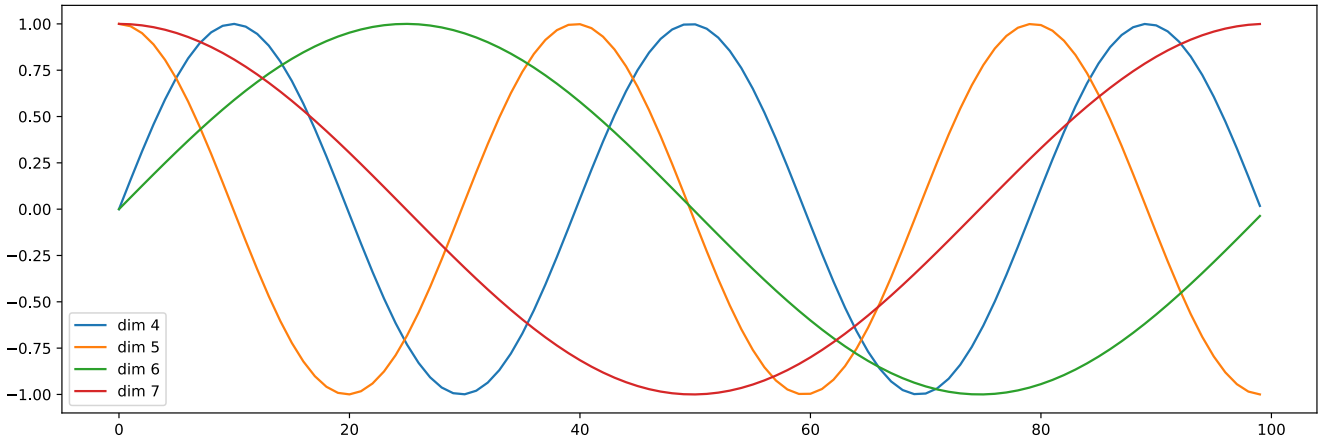
位置符号器 Position encoders

トランスフォーマーの入力には，上述の単語表現に加えて，位置符号器からの信号も重ね合わされる。 位置 iの信号は次式で周波数領域へと変換される:

$$\mathrm{PE}_{(\mathrm{pos},2i)} = \sin\left(\frac{\mathrm{pos}}{10000^{\frac{2i}{d_{\mathrm{model}}}}}\right) \tag{5}$$

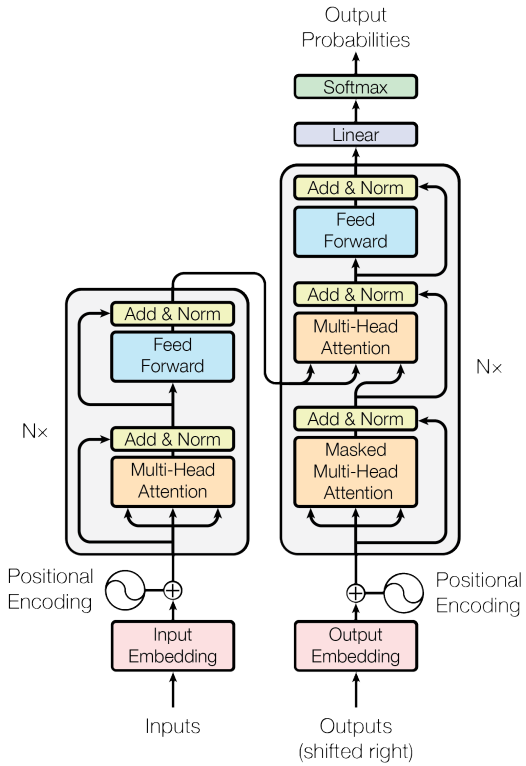
$$\mathrm{PE}_{(\mathrm{pos},2i+1)} = \cos\left(\frac{\mathrm{pos}}{10000^{\frac{2i}{d_{\mathrm{model}}}}}\right) \tag{6}$$

位置符号器による位置表現は，i番目の位置情報をワンホット表現するのではなく，周波数領域に変換することで周期情報を表現する試みと見なし得るだろう。



位置符号化に用いられる符号化

このようにしてできた値を入力側と出力側で下図のように連結させたものがトランスフォーマーである。



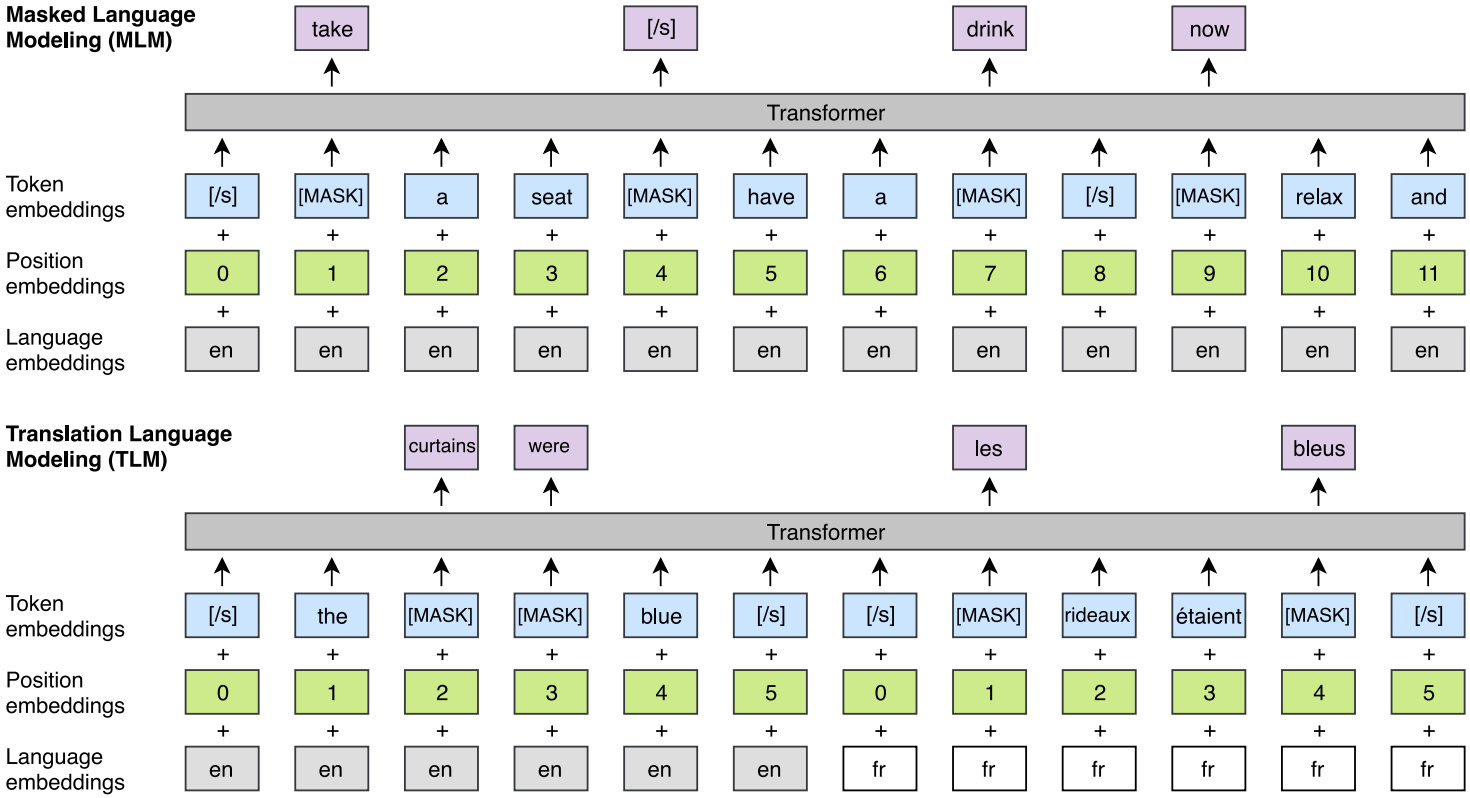
From (Vaswani et al. 2017) Fig. 1

これまで見てきたように，トランスフォーマーでは入力信号に基づいて情報の変換が行なわれる。この意味ではトランスフォーマーにおける 多頭 自己注意 MHSA とはボトムアップ注意の変形であるとみなしうる。 逆言すれば，RNN のように過去の履歴をすべて保持しているわけではないので，系列情報については，position encoders に頼っている側面が指摘できる。

BERT, GPT, ELMo 事前訓練の違い

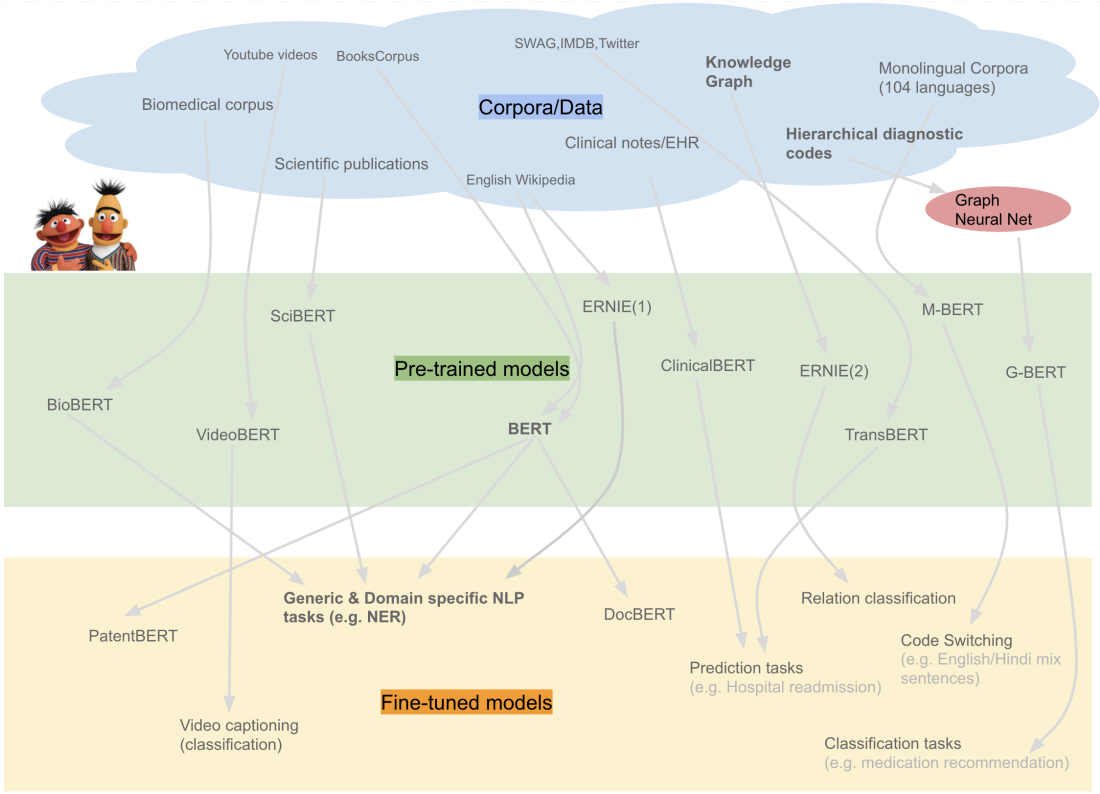
- BERT: トランスフォーマー，マスク化言語モデル，次文予測課題
- GPT: 順方向トランスフォーマー
- ELMo: 双方向 RNN による中間層の連結

多言語対応



From (Lample and Conneau 2019) Fig. 1

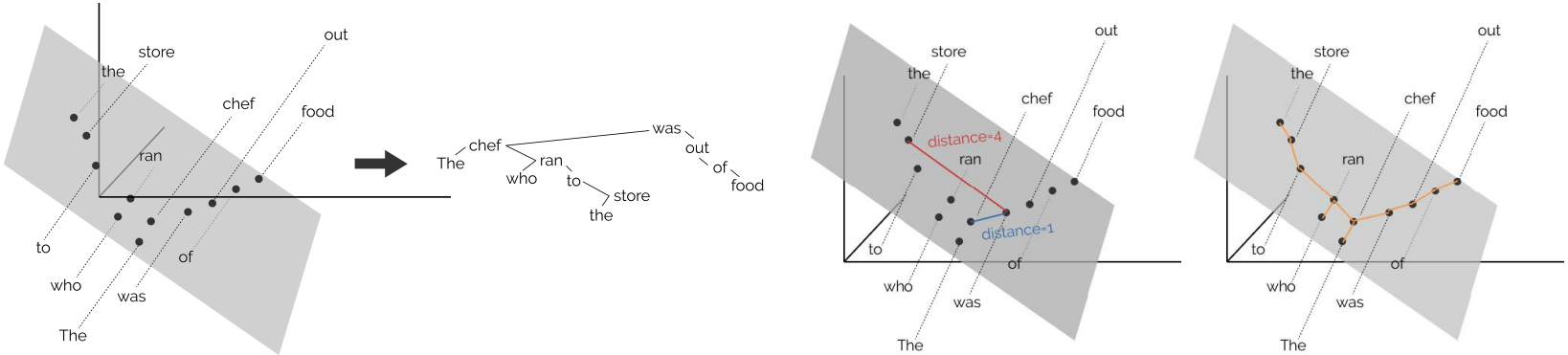
BERT の発展



From https://towardsdatascience.com/a-review-of-bert-based-models-4ffdc0f15d58

BERT: 埋め込みモデルによる構文解析

BERT の構文解析能力を下図示した。 各単語の共通空間に射影し， 単語間の距離を計算することにより構文解析木と同等の表現を得ることができることが報告されている(Hewitt and Manning 2019)。



BERT による構文解析木を再現する射影空間

From https://github.com/john-hewitt/structural-probes

word2vec において単語間の距離は内積で定義されていた。 このことから， 文章を構成する単語で張られる線形内積空間内の距離が構文解析木を与えると見なすことは不自然ではない。 そこで構文解析木を再現するような射影変換を見つけることができれば BERT を用いて構文解析が可能となる。 例えば上図における chef と store と was の距離を解析木を反映するような空間を見つけ出すことに相当する。 2 つの単語 w_i, w_j し単語間の距離を $d(w_i, w_j)$ とする。適当な変換を施した後の座標を h_i, h_j とすれば， 求める変換 B は次式のような変換を行なうことに相当する:

$$\min_{B} \sum_{l} \frac{1}{|s_\ell|^2} \sum_{i,j} \left(d\left(w_i, w_j\right) - \left\| B\left(h_i - h_j\right) \right\|^2 \right) \quad (7$$

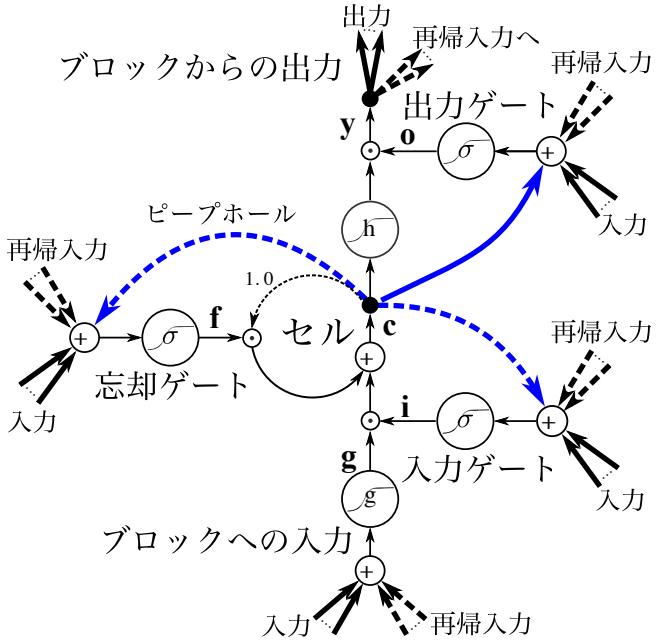
ここで d は文 s の訓練文のインデックスであり，各文の長さで規格化することを意味している。

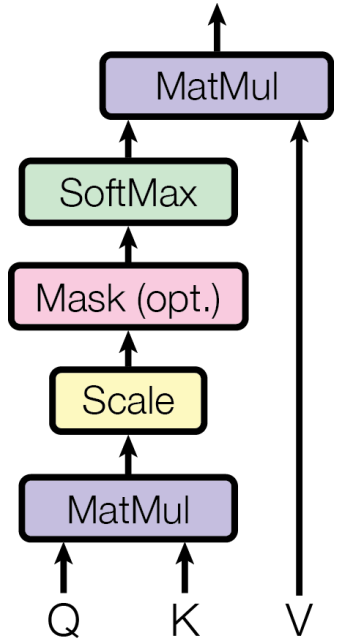
BERT 実装

BERT 実装のパラメータを以下に示した。 現在配布されている BERT-base あるいは性能が良い BERT-large は各層のニューロン数と全体の層数である。

- データ: Wikipedia (2.5B words) + BookCorpus (800M words)
- バッチサイズ: 131,072 words (1024 sequences × 128 length or 256 sequences × 512 length)
- 訓練ステップ: 1M steps (40 epochs)
- 最適化アルゴリズム: AdamW, 1e-4 learning rate, linear decay
- BERT-Base: 12 層, 各層 768 ニューロン, 12 多頭注意
- BERT-Large: 24 層, 各層 1024 ニューロン, 16 多頭注意
- 訓練時間: 4x4 / 8x8 の TPU で 4 日間

LSTM





左: LSTM (浅川, 2015) より，右: トランスフォーマー(Vaswani et al. 2017)
入力ゲートと入力 は Q, K と同一視，出力ゲートと V とは同一視可能？

BERT embeddings

```
class BertEmbeddings(nn.Module):
    """Construct the embeddings from word, position and token_type embeddings.
    """

    def __init__(self, config):
        super().__init__()
        self.word_embeddings = nn.Embedding(config.vocab_size, config.hidden_size,
padding_idx=config.pad_token_id)
        self.position_embeddings = nn.Embedding(config.max_position_embeddings, config.hidden_size)
        self.token_type_embeddings = nn.Embedding(config.type_vocab_size, config.hidden_size)

        # self.LayerNorm is not snake-cased to stick with TensorFlow model variable name and be able to load
        # any TensorFlow checkpoint file
        self.LayerNorm = BertLayerNorm(config.hidden_size, eps=config.layer_norm_eps)
        self.dropout = nn.Dropout(config.hidden_dropout_prob)

    def forward(self, input_ids=None, token_type_ids=None, position_ids=None, inputs_embeds=None):
        if input_ids is not None:
            input_shape = input_ids.size()
        else:
            input_shape = inputs_embeds.size()[:-1]

        seq_length = input_shape[1]
        device = input_ids.device if input_ids is not None else inputs_embeds.device
        if position_ids is None:
            position_ids = torch.arange(seq_length, dtype=torch.long, device=device)
            position_ids = position_ids.unsqueeze(0).expand(input_shape)
        if token_type_ids is None:
            token_type_ids = torch.zeros(input_shape, dtype=torch.long, device=device)

        if inputs_embeds is None:
            inputs_embeds = self.word_embeddings(input_ids)
        position_embeddings = self.position_embeddings(position_ids)
        token_type_embeddings = self.token_type_embeddings(token_type_ids)

        embeddings = inputs_embeds + position_embeddings + token_type_embeddings
        embeddings = self.LayerNorm(embeddings)
        embeddings = self.dropout(embeddings)
        return embeddings
```

BERT inside

```
        query_layer = self.transpose_for_scores(mixed_query_layer)
        key_layer = self.transpose_for_scores(mixed_key_layer)
        value_layer = self.transpose_for_scores(mixed_value_layer)

        # Take the dot product between "query" and "key" to get the raw attention scores.
        attention_scores = torch.matmul(query_layer, key_layer.transpose(-1, -2))
        attention_scores = attention_scores / math.sqrt(self.attention_head_size)
        if attention_mask is not None:
            # Apply the attention mask is (precomputed for all layers in BertModel forward() function)
            attention_scores = attention_scores + attention_mask

        # Normalize the attention scores to probabilities.
        attention_probs = nn.Softmax(dim=-1)(attention_scores)

        # This is actually dropping out entire tokens to attend to, which might
        # seem a bit unusual, but is taken from the original Transformer paper.
        attention_probs = self.dropout(attention_probs)

        # Mask heads if we want to
        if head_mask is not None:
            attention_probs = attention_probs * head_mask

        context_layer = torch.matmul(attention_probs, value_layer)
```

第 3 部

流行りの句

流行りの句

Published as a conference paper at ICLR 2020

REFORMER: THE EFFICIENT TRANSFORMER

Nikita Kitaev* U.C. Berkeley & Google Research kitaev@cs.berkeley.edu
Łukasz Kaiser* Google Research {lukaszkaiser,levskaya}@google.com
Anselm Levskaya Google Research

ABSTRACT

Large Transformer models routinely achieve state-of-the-art results on a number of tasks but training these models can be prohibitively costly, especially on long sequences. We introduce two techniques to improve the efficiency of Transformers. For one, we replace dot-product attention by one that uses locality-sensitive hashing, changing its complexity from $O(L^2)$ to $O(L \log L)$, where L is the length of the sequence. Furthermore, we use reversible residual layers instead of the standard residuals, which allows storing activations only once in the training process instead of N times, where N is the number of layers. The resulting model, the Reformer, performs on par with Transformer models while being much more memory-efficient and much faster on long sequences.

Published as a conference paper at ICLR 2020

ON THE RELATIONSHIP BETWEEN SELF-ATTENTION AND CONVOLUTIONAL LAYERS

Jean-Baptiste Cordonnier, Andreas Loukas & Martin Jaggi
École Polytechnique Fédérale de Lausanne (EPFL)
{first.last}@epfl.ch

ABSTRACT

Recent trends of incorporating attention mechanisms in vision have led researchers to reconsider the supremacy of convolutional layers as a primary building block. Beyond helping CNNs to handle long-range dependencies, Ramachandran et al. (2019) showed that attention can completely replace convolution and achieve state-of-the-art performance on vision tasks. This raises the question: do learned attention layers operate similarly to convolutional layers? This work provides evidence that attention layers can perform convolution and, indeed, they often learn to do so in practice. Specifically, we prove that a multi-head self-attention layer with sufficient number of heads is at least as expressive as any convolutional layer. Our numerical experiments then show that self-attention layers attend to pixel-grid patterns similarly to CNN layers, corroborating our analysis. Our code is publicly available[1].

流行りの句 (continued.)

NeurIPS 2019 Workshop on Machine Learning with Guarantees, Vancouver, Canada.

Are Transformers universal approximators of sequence-to-sequence functions?

Chulhee Yun*
MIT
chulheey@mit.edu

Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, Sanjiv Kumar
Google Research New York
{bsrinadh,ankitsrawat,sashank,sanjivk}@google.com

Abstract

Despite the widespread adoption of Transformer models for NLP tasks, the expressive power of these models is not well-understood. In this paper, we establish that Transformer models are universal approximators of continuous *permutation equivariant* sequence-to-sequence functions with compact support, which is quite surprising given the amount of shared parameters in these models. Furthermore, using positional encodings, we circumvent the restriction of permutation equivariance, and show that Transformer models can universally approximate *arbitrary* continuous sequence-to-sequence functions on a compact domain. Interestingly, our proof techniques clearly highlight the different roles of the self-attention and the feed-forward layers in Transformers. In particular, we prove that fixed width self-attention layers can compute *contextual mappings* of the input sequences, playing a key role in the universal approximation property of Transformers. Based on this insight from our analysis, we consider other simpler alternatives to self-attention layers and empirically evaluate them.

The 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019

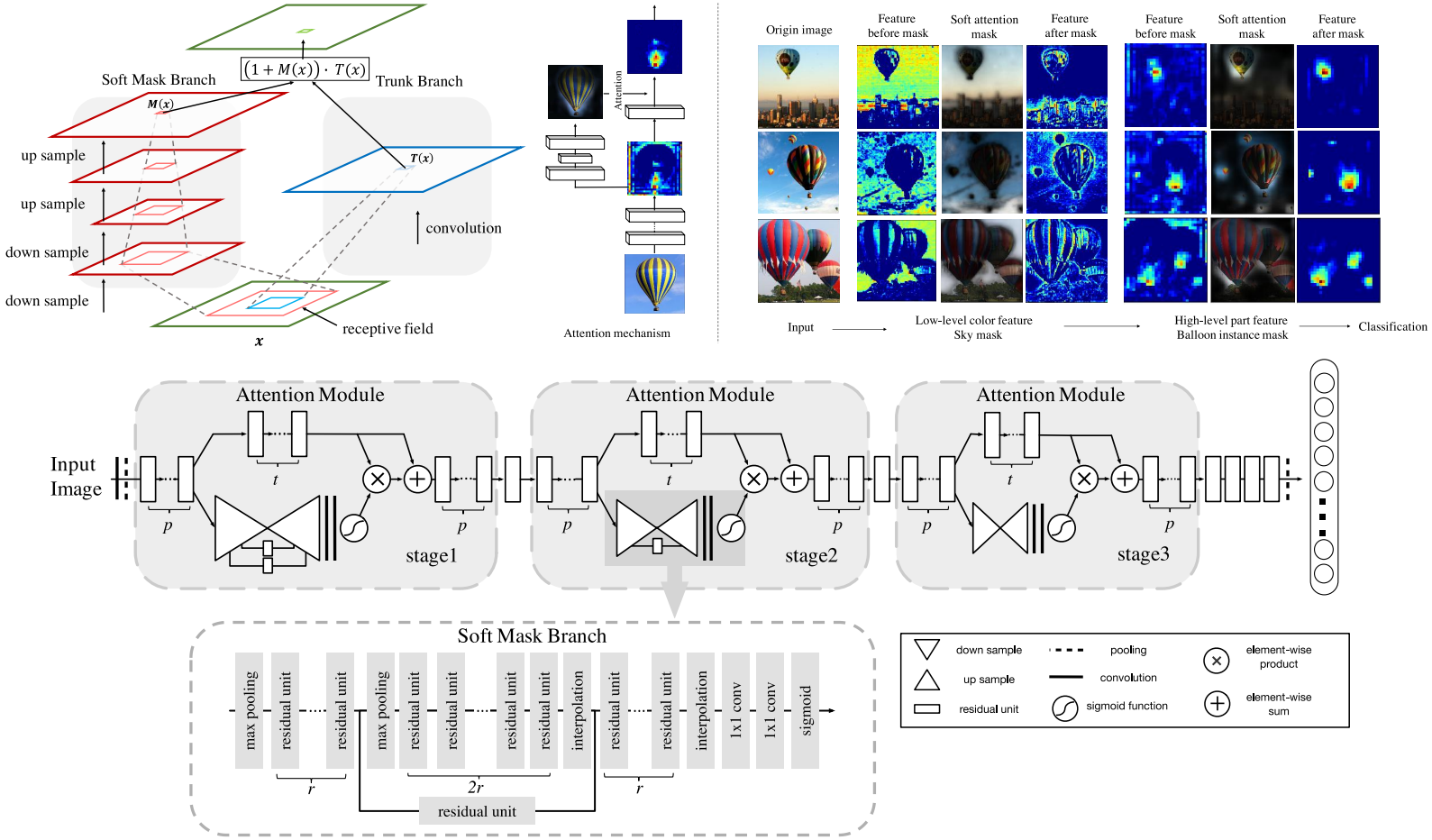
DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter

Victor SANH, Lysandre DEBUT, Julien CHAUMOND, Thomas WOLF
Hugging Face
{victor,lysandre,julien,thomas}@huggingface.co

Abstract

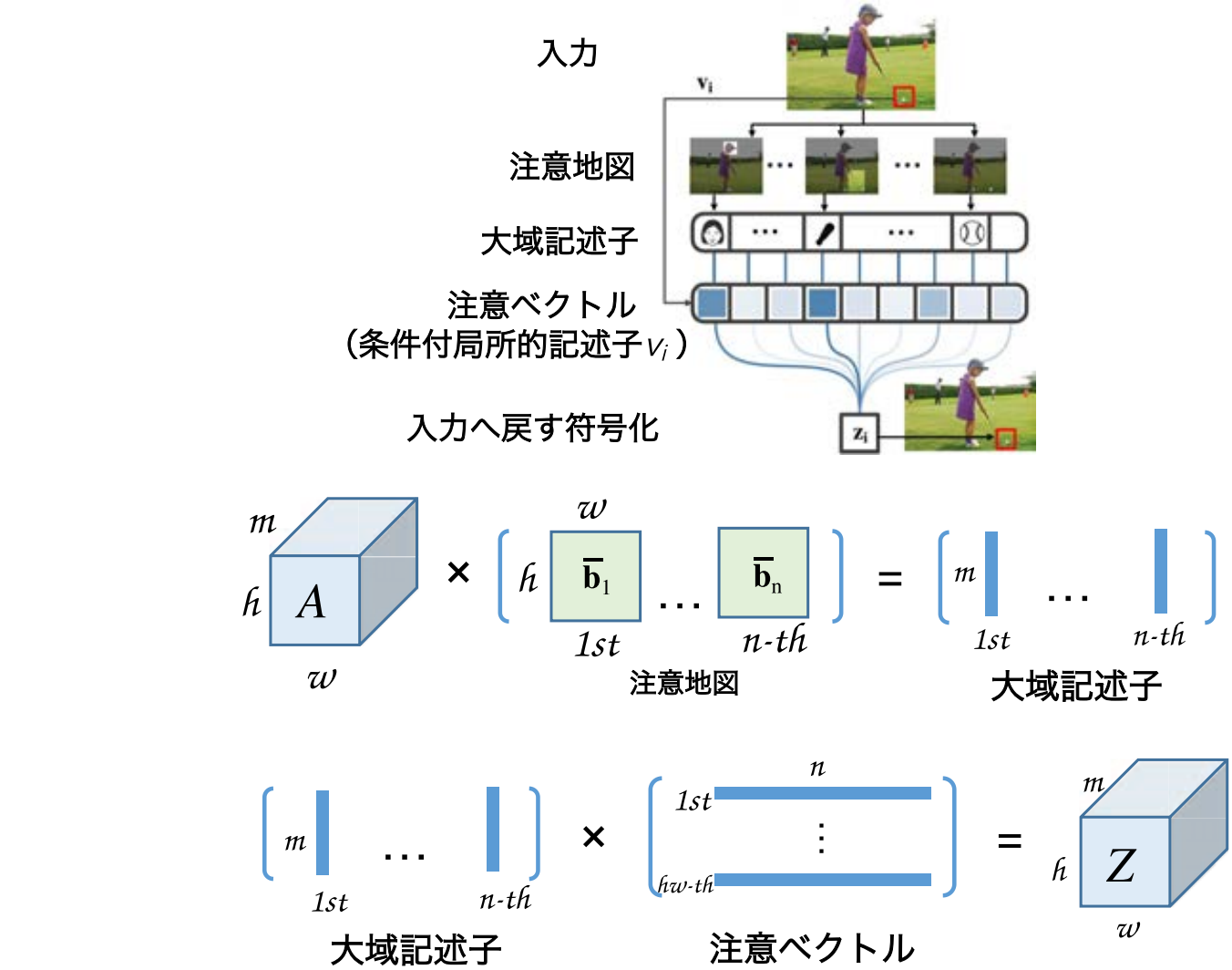
As Transfer Learning from large-scale pre-trained models becomes more prevalent in Natural Language Processing (NLP), operating these large models in on-the-edge and/or under constrained computational training or inference budgets remains challenging. In this work, we propose a method to pre-train a smaller general-purpose language representation model, called DistilBERT, which can then be fine-tuned with good performances on a wide range of tasks like its larger counterparts. While most prior work investigated the use of distillation for building task-specific models, we leverage knowledge distillation during the pre-training phase and show that it is possible to reduce the size of a BERT model by 40%, while retaining 97% of its language understanding capabilities and being 60% faster. To leverage the inductive biases learned by larger models during pre-training, we introduce a triple loss combining language modeling, distillation and cosine-distance losses. Our smaller, faster and lighter model is cheaper to pre-train and we demonstrate its capabilities for on-device computations in a proof-of-concept experiment and a comparative on-device study.

Residual attention



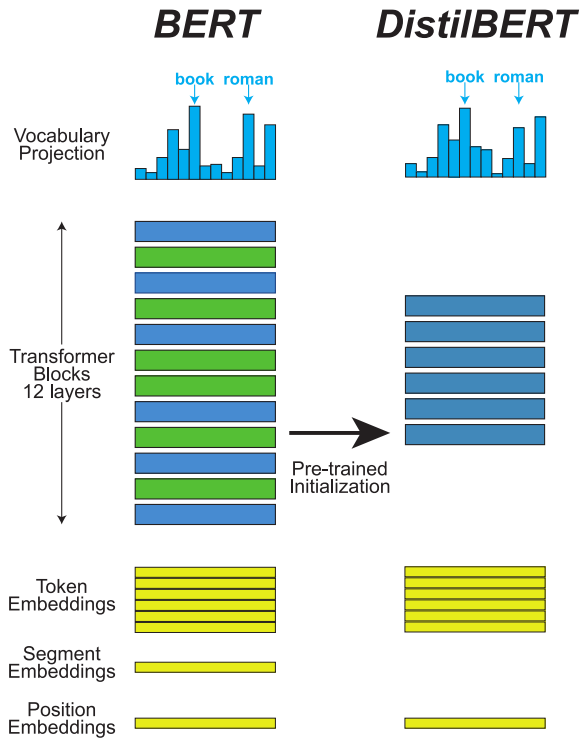
(Wang et al. 2017) Fig. 1, 2, 3

A2 net



From (Chen et al. 2018) Fig. 1

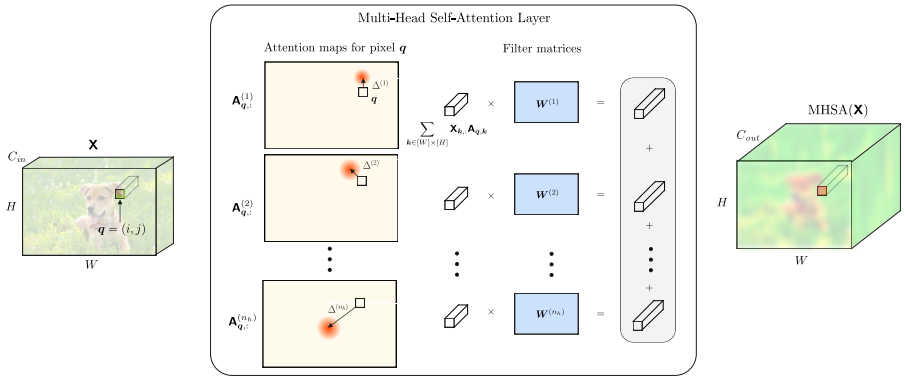
DistilBERT



3 つの損失関数(Sanh et al. 2020):

1. 知識蒸留損失
2. マスク化言語モデル損失
3. コサイン損失

Relationship between self-attention and convolution



C Positional Encoding References

Model	type of positional encoding			relative
	sinusoids	learned	quadratic	
Vaswani et al. (2017)	✓			
Radford et al. (2018)		✓		
Devlin et al. (2018)		✓		
Dai et al. (2019)	✓			✓
Yang et al. (2019)	✓			✓
Bello et al. (2019)		✓		✓
Ramachandran et al. (2019)		✓		✓
Our work		✓	✓	✓

Table 3: Types of positional encoding used by transformers models applied to text (*top*) and images (*bottom*). When multiple encoding types have been tried, we report the one advised by the authors.

From (Cordonnier, Loukas, and Jaggi 2020)

第 3 部まとめ

- MHSA は 畳み込み と同等の能力がありそうである。
- Reformer に見られるように position encodings を工夫する余地は残されているように思われる。

第 4 部

不易の句

Dicotomy

- ボトムアップ と トップダウン
- 何 と 何処 (腹側 背側)
- 特徴，対象，場所へ向けられるの注意
- 外発的，内発的 注意

関連脳領域

- FEF 前頭眼野 (Monosov and Thompson 2009)
- Lateral Intraparietal area (LIP) 側頭頭頂領域 (Wardak, Olivier, and Duhamel 2004)
- Superior Colliculus(SC) 上丘 (Krauzlis, Lovejoy, and Zénon 2013)
- PFC 前頭皮質 (Miller and Cohen 2001)
- VPA (Bichot et al. 2015)

認知心理学分野

- フィルタリング [Broadbent (1958)}, 減衰説 (Treisman 1969)
- 特徴統合理論 (Treisman and Gelade 1980);(Treisman 1988)
- Guided Search 2.0 (Wolfe 1994)
- 目標／妨害刺激類似性: (Duncan and Humphreys 1989, 1992DuncanHumphreys_engagement)
- サーチライト(スポットライト)仮説 (Crick 1984), ズームレンズ(Eriksen and St.James 1986)
- 勝者占有回路(Koch and Ullman 1985) = softmax

計算モデル (Implementation)

- (Milanese et al. 1994)
- (Itti, Koch, and Niebur 1998)
- (Borji and Itti 2013) SOTA

総説論文

- (Itti and Koch 2001)
- (Knudsen 2007)
- (Petersen and Posner 2012)
- (Kimura, Yonetani, and Hirayama 2013)
- (Itti and Borji 2015) Oxford Handbook of attention

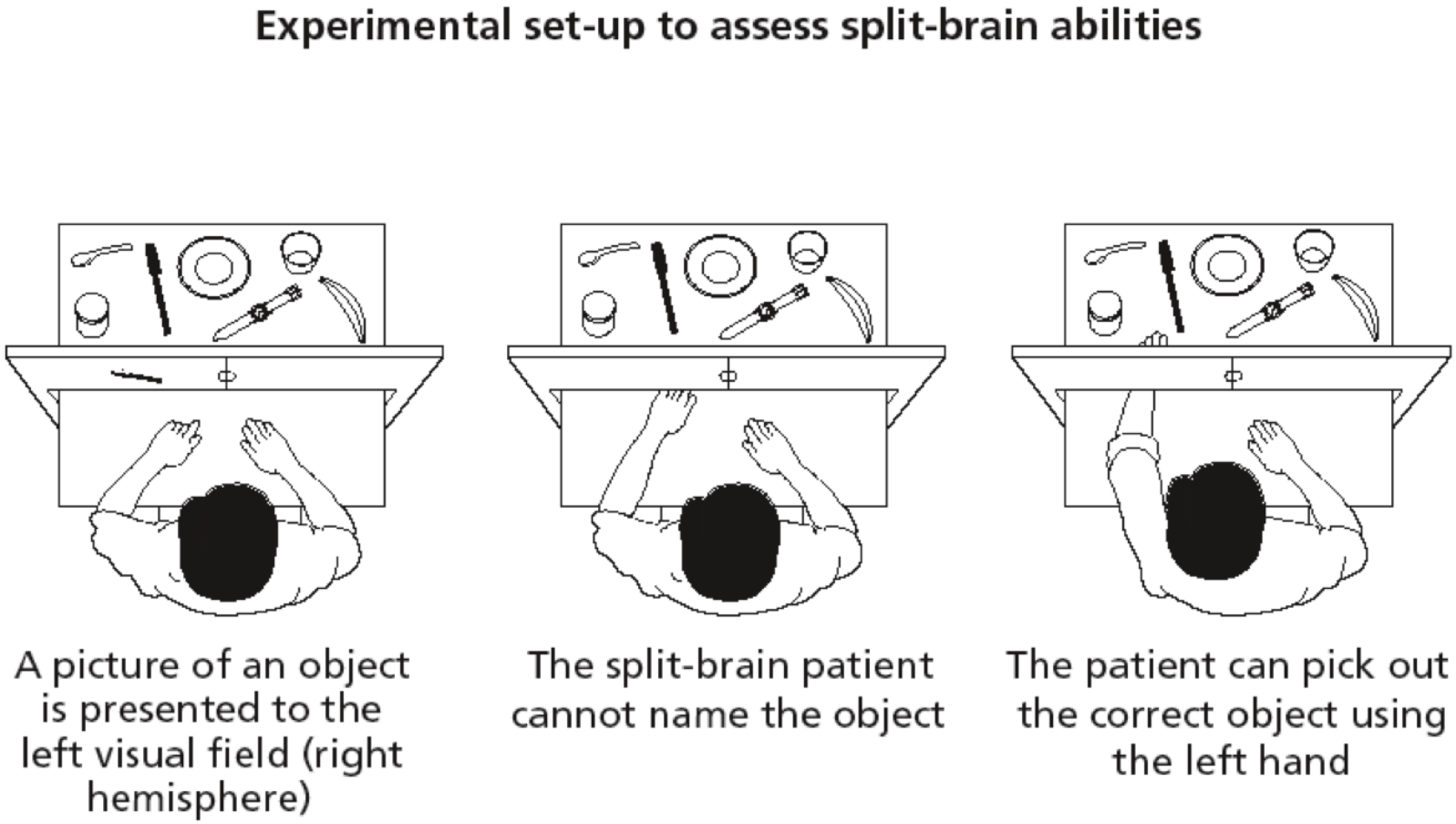
深層学習系

- 自動翻訳 (Bahdanau, Cho, and Bengio 2015, 2015Luong_attention)
- 画像脚注付け (Vinyals et al. 2015)
- 注意 (W. Wang and Shen 2018)

温故知新

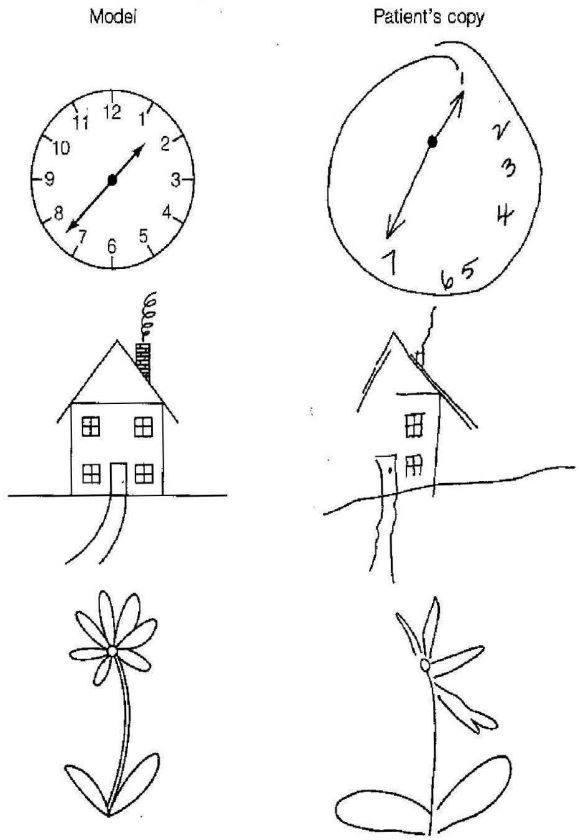
- 脳梁切断患者による分離脳 (Sperry 1961)
- 半側空間無視 (Heilman and Valenstein 1979)
- 頭頂葉損傷患者の注意のディスエンゲージメント(Posner 1980)
- 両耳分離聴実験, カクテルパーティ効果 (Broadbent 1958);(Treisman 1964)
- 特徴統合理論[Treisman and Gelade (1980),1988Treisman}
- 計算論的モデル サーチライト(スポットライト)仮説 (Crick 1984)
- モデルとデータセット公開, 競技会 (Itti and Koch 2001);(Itti and Borji 2014)
- DeepGazeII (Kümmerer et al. 2017)

分離脳 Split brain



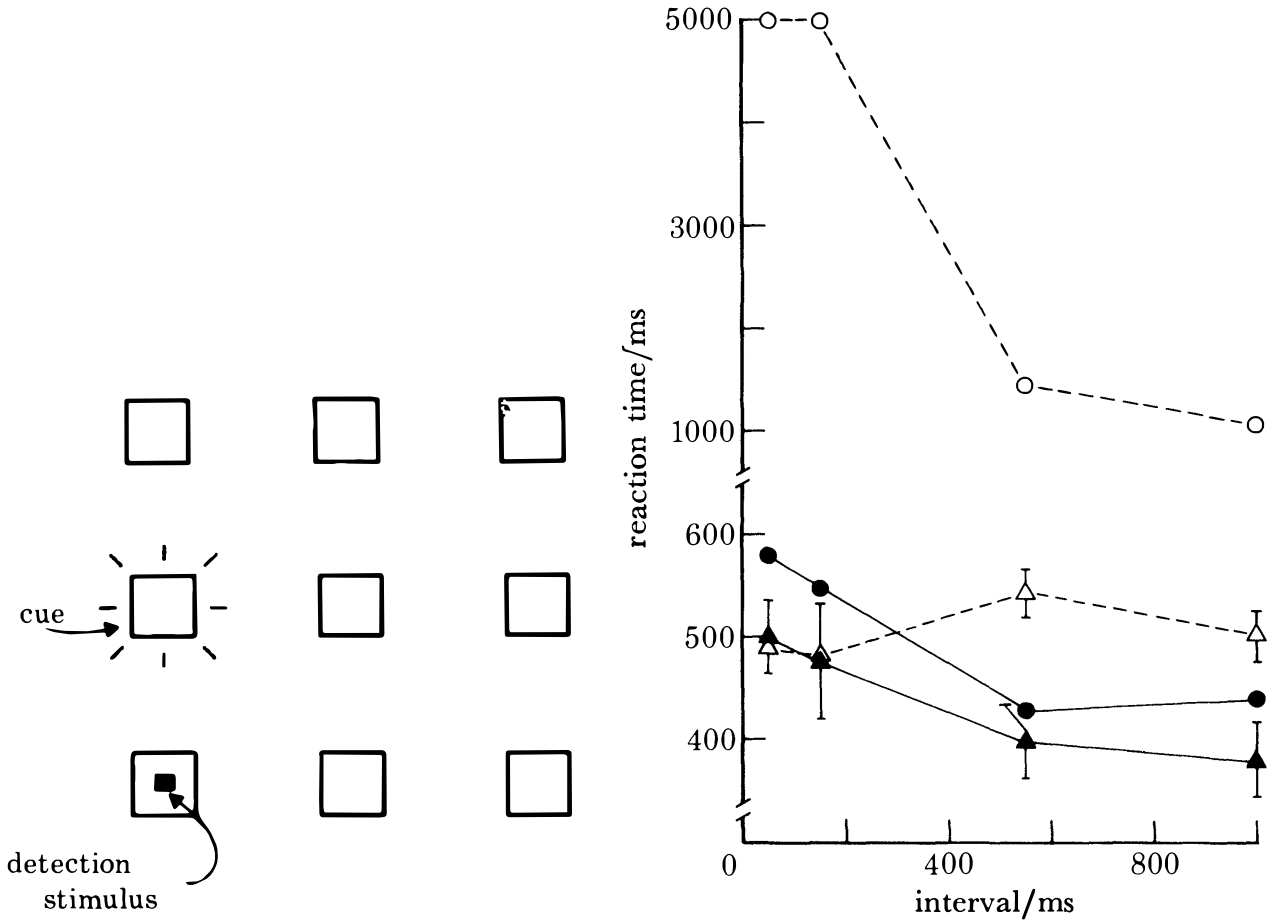
From (Sperry 1968) Fig. 5

半側空間無視



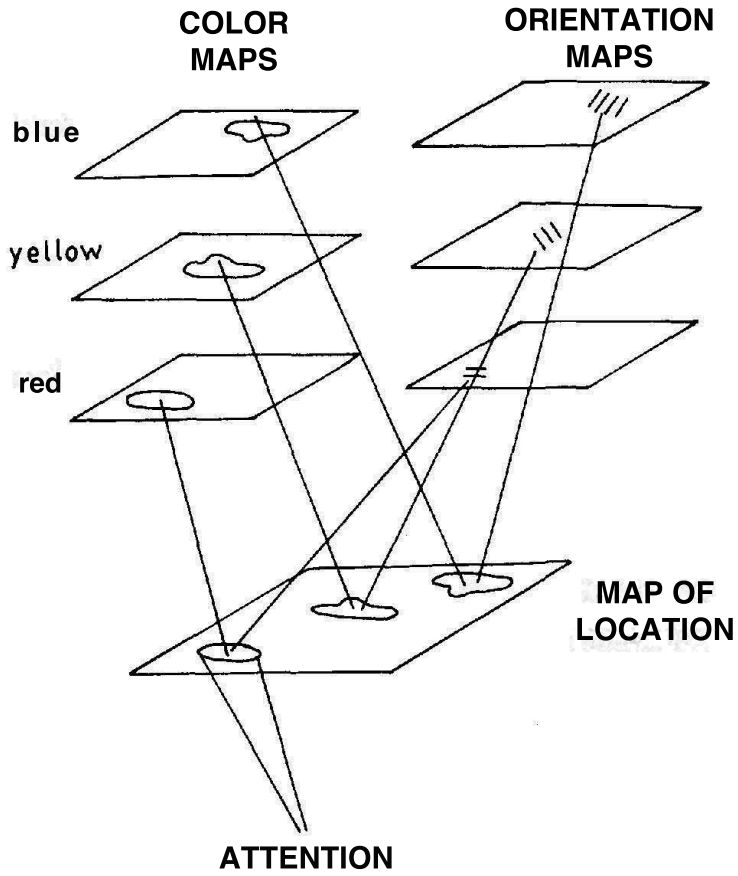
From (Bloom and Lazerson 1988) Fig. 17-6

ポズナーとコーヘン



From (Posner 1980) Fig. 1, Fig.6: 右頭頂葉障害を呈した患者 (R.S.) の結果。 円:ターゲットが左視野提示, 三角:ターゲット右視野提示。 白点線:非有効手がかり, 黒実線:有効手がかり。横軸は ISI。縦軸は反応時間中央値

特徴統合理論 (FIT)



From (Treisman and Souther 1985) Fig. 9

探索非対称性 search asymmetry}

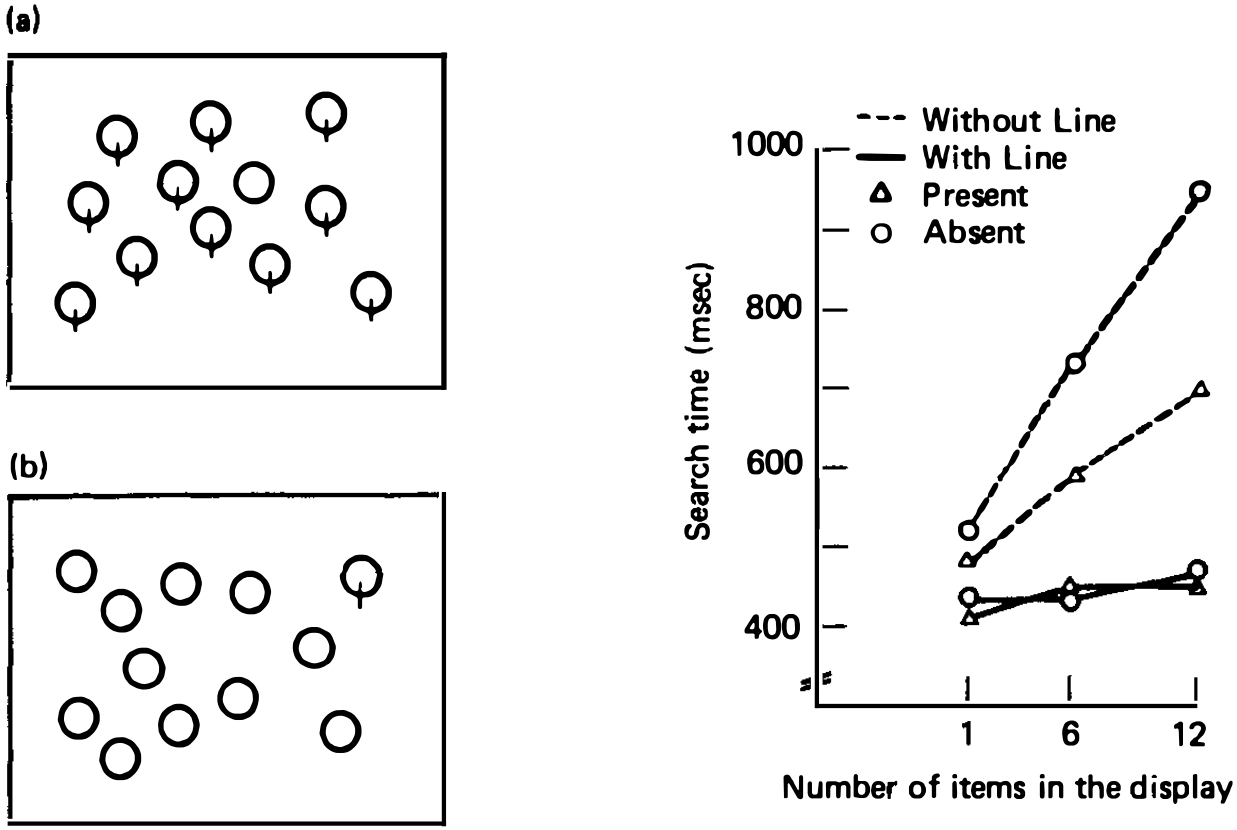


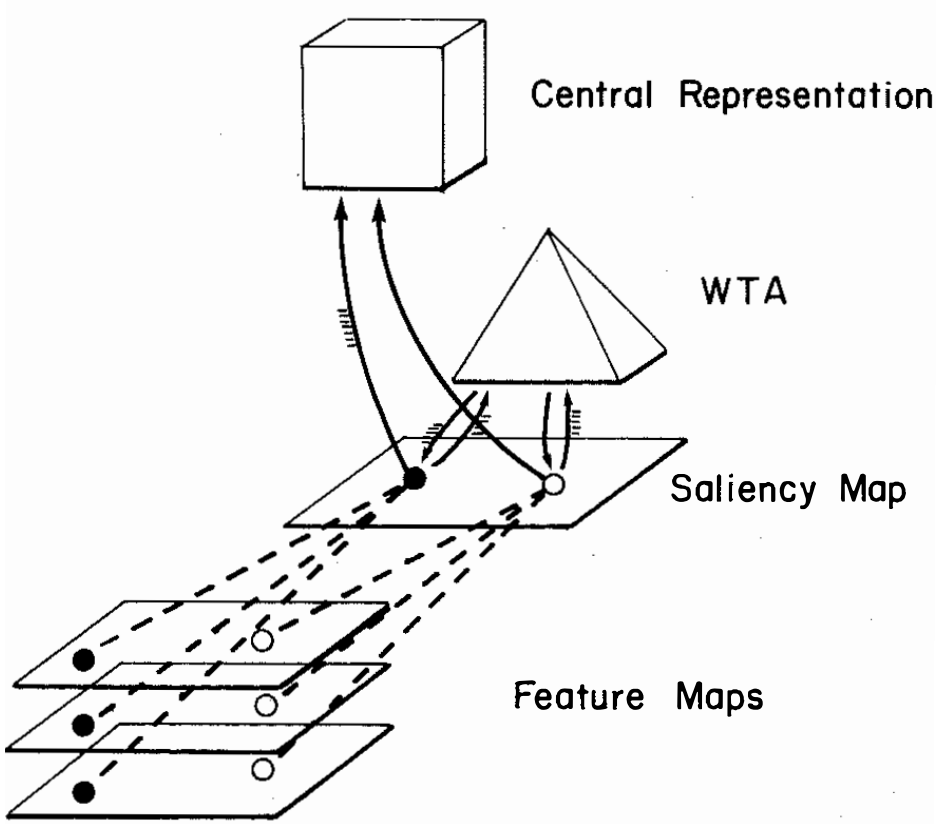
FIG. 3. Examples of displays and mean search times for a target circle with and without an intersecting line.

From [Treisman (1988)} Fig. 3

上図右の結果は横軸に同時に提示された刺激の個数であり，縦軸は反応時間です。 線分特徴が存在する刺激 (Q) が目標となるか，存在しない (O) が目標となるか によって反応時間に差が認められます。結果は点線，すあんわち特徴が存在しな い目標を探索する条件，点線で描画，では同時に提示された刺激数が増加するに 従って反応時間が増大します。一方，特徴が存在する目標を探索する条

件では, 同時提示された刺激の個数によらず反応時間は平坦になります。 以下に同様な実験結果を示しました。

スポットライトメタファー



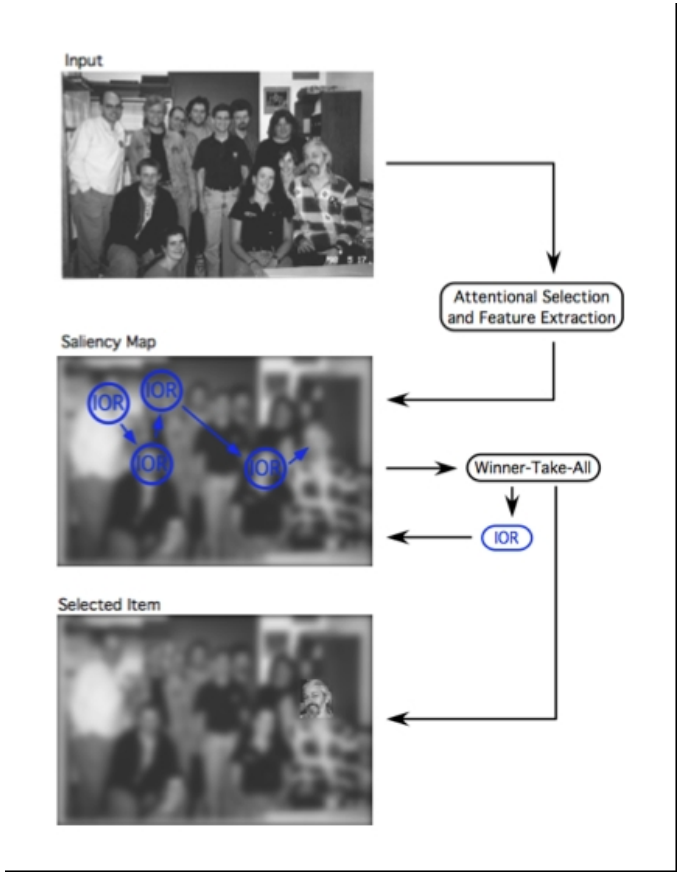
From (Koch and Ullman 1985) Fig. 5

- スポットライトメタファー (Crick 1984)

  Attention can be likened to a spotlight that enhances the efficiency of detection of events within its beam. Unlike when acuity is involved, the effeet of the beam is not related to the fovea. When the fovea is unill uminated by attention, its ability to lead to detection is diminished, as would be the case with any other are a of the visual system. Posner p.172

- (Summerfield et al. 2006) は AI の研究にも影響
- ネットワークの内部メモリから読み出す情報を選択するために注意機構
- 機械翻訳 (Bahdanau, Cho, and Bengio 2015), NTM (Graves et al. 2016)
- コンテンツアドレス(Hopfield 1982)
- BERT (Devlin et al. 2018)

Inhibition of Return (IOR)



From http://www.scholarpedia.org/article/Inhibition_of_return

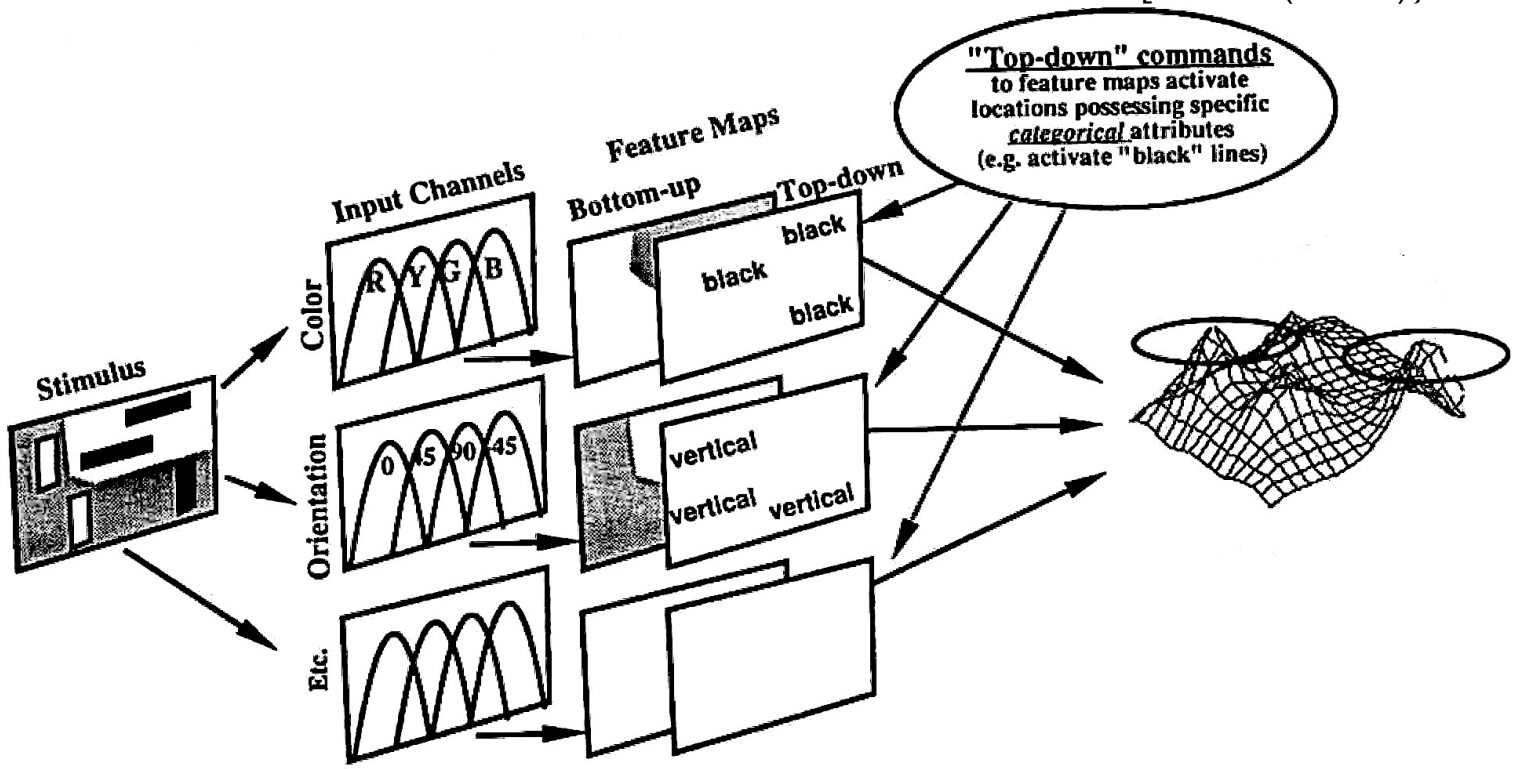
From The superior colliculus (SC) has been implicated as the neural substrate for IOR through four converging, but indirect, lines of evidence.

1. IOR is abnormal in patients with midbrain degeneration due to progressive supranuclear palsy (PSP).

2. It is preserved in patients with hemianopia, a condition in which only extrageniculate pathways are available to process visual information.
3. It is present in newborn infants, in whom the geniculostriate pathways are not yet developed.
4. It is generated asymmetrically in temporal and nasal visual fields, suggesting retinotecal mediation.

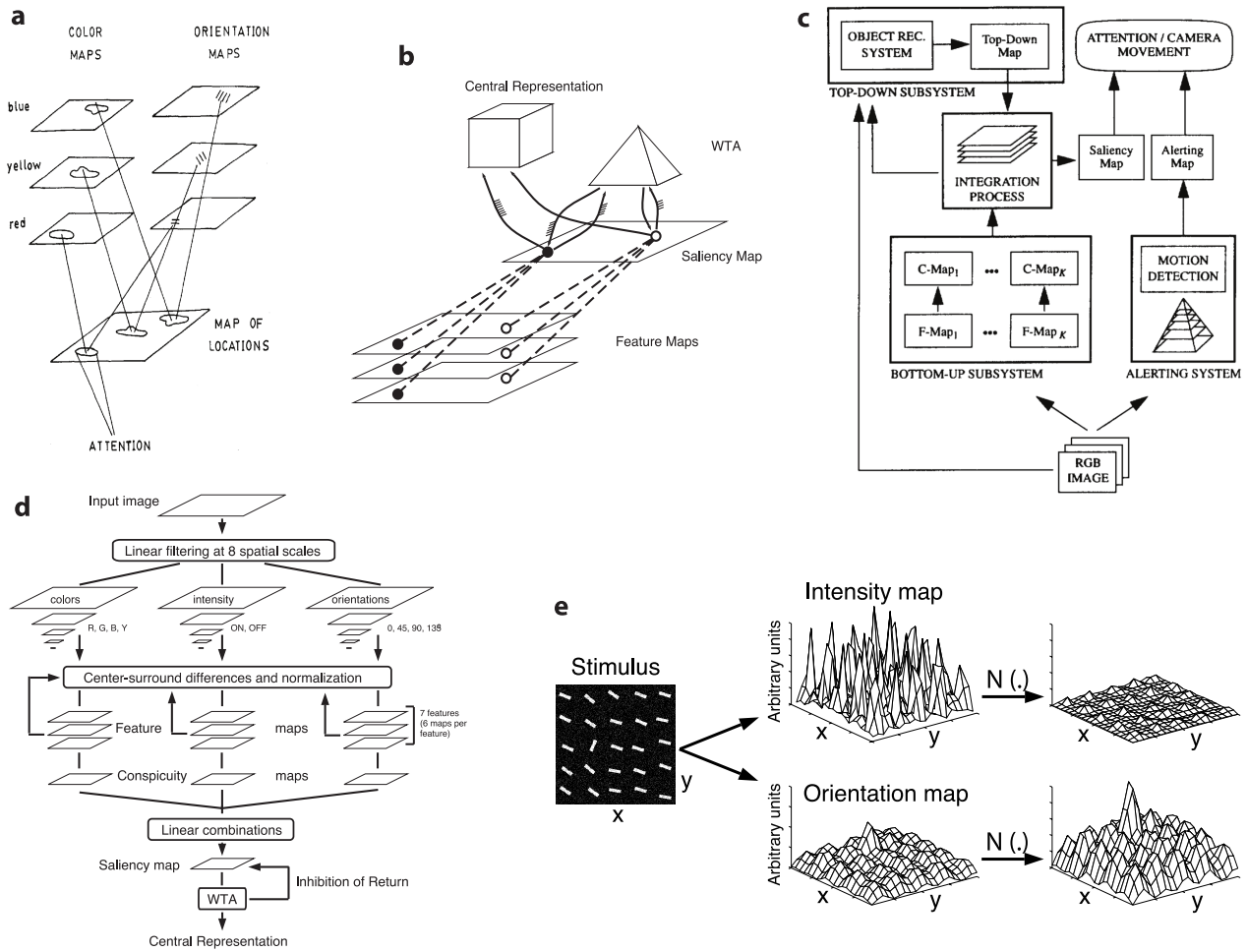
ガイド付き探索モデル Guided Search 2.0

最初にトップダウン注意を明示的に示した *ガイド付き探索モデル** [Wolfe (1994)}



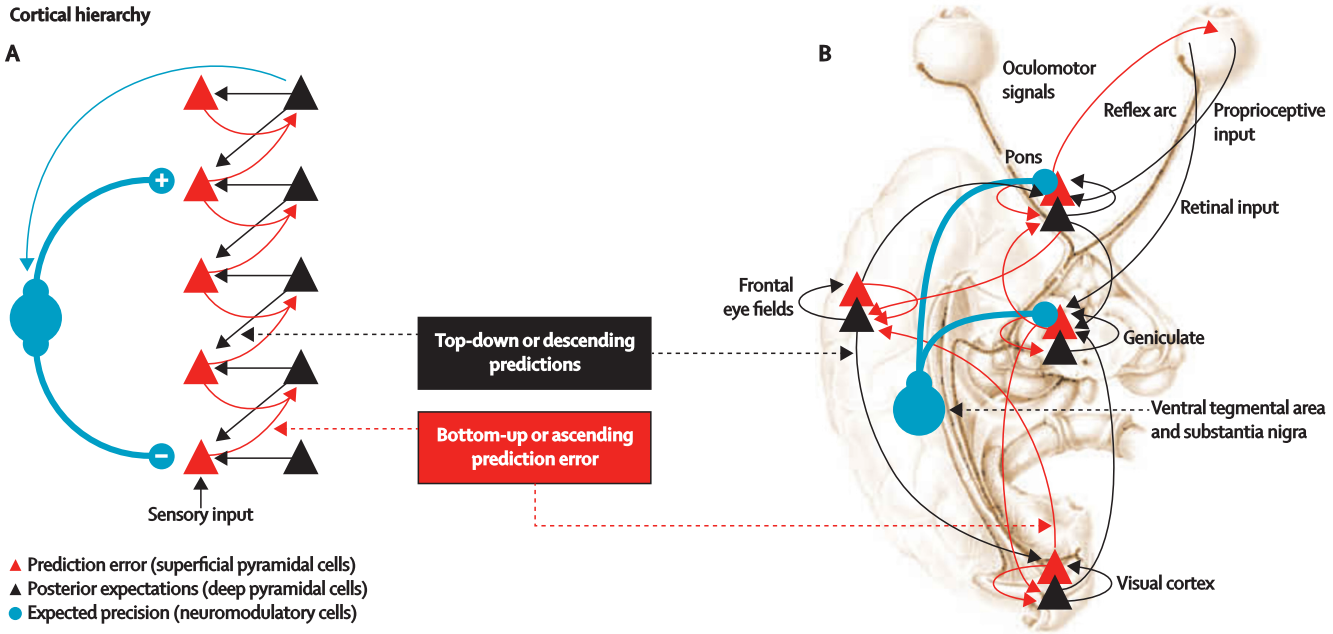
From (Wolfe 1994) Fig. 2

(Itti and Borji 2015) の総説論文からそれまでのモデルの概説図



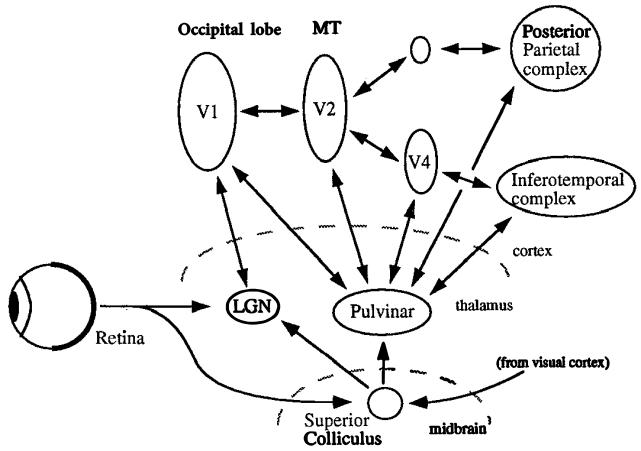
From (Itti and Borji 2015) Fig. 2

Friston's attetion



From (Friston et al. 2014) Fig. 1

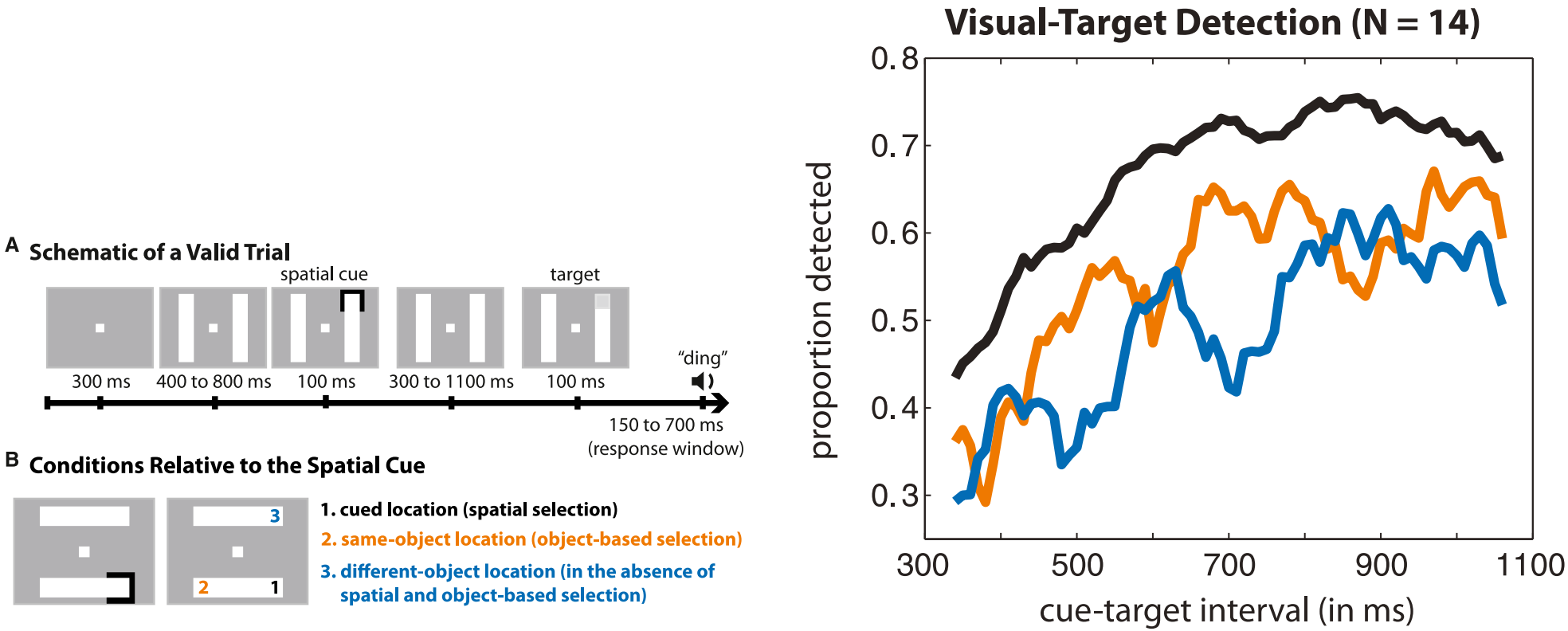
上丘 SC



From (Olshausen, Anderson, and Essen 1993) Fig. 10a

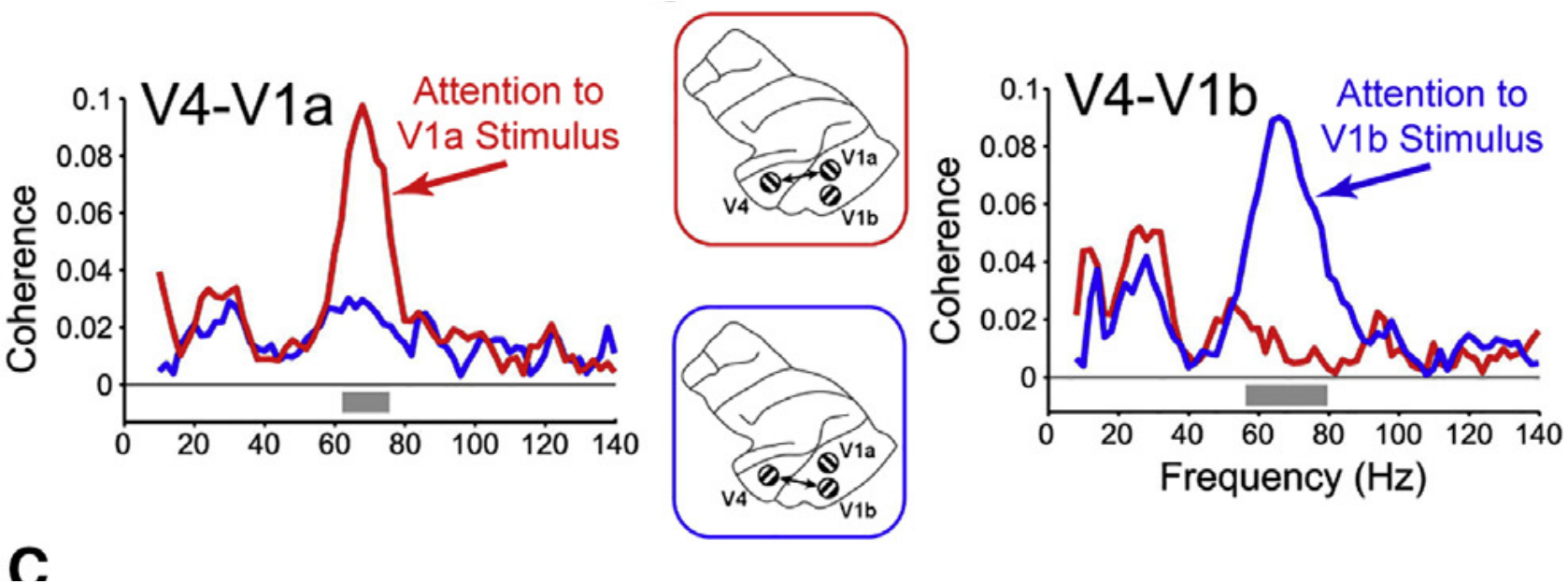
- 霊長類の視覚系の動作は注意を伴う視線の移動により外界を認識
- すべての入力を並行して処理するのではなく，視覚的注意は場所や物体間の遷移(Koch and Ullman 1985; Moore and Zirnsak 2017; Posner and Petersen 1990)
- 情報の優先順位付け，取捨選択(Olshausen, Anderson, and Essen 1993; Salinas and Abbott 1997

リズム現象



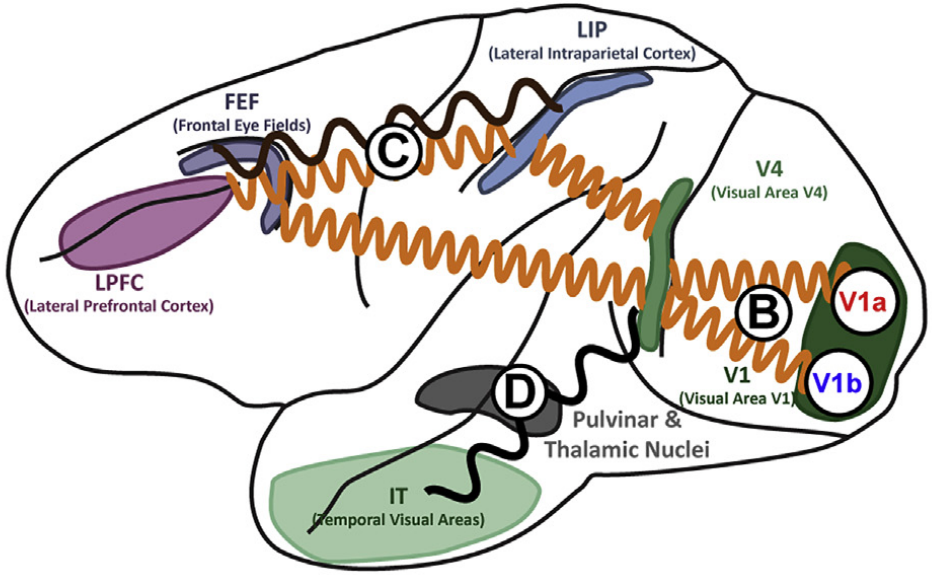
From (Fiebelkorn, Saalmann, and Kastner 2013) Fig. 1 and Fig. 2a

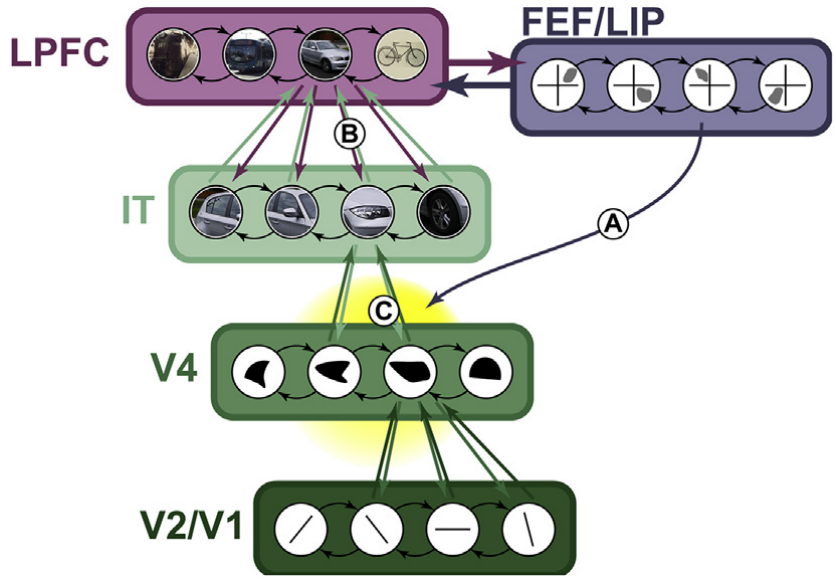
リズム現象 (2)



From (Buschman and Kastner 2015) Fig. 3b}

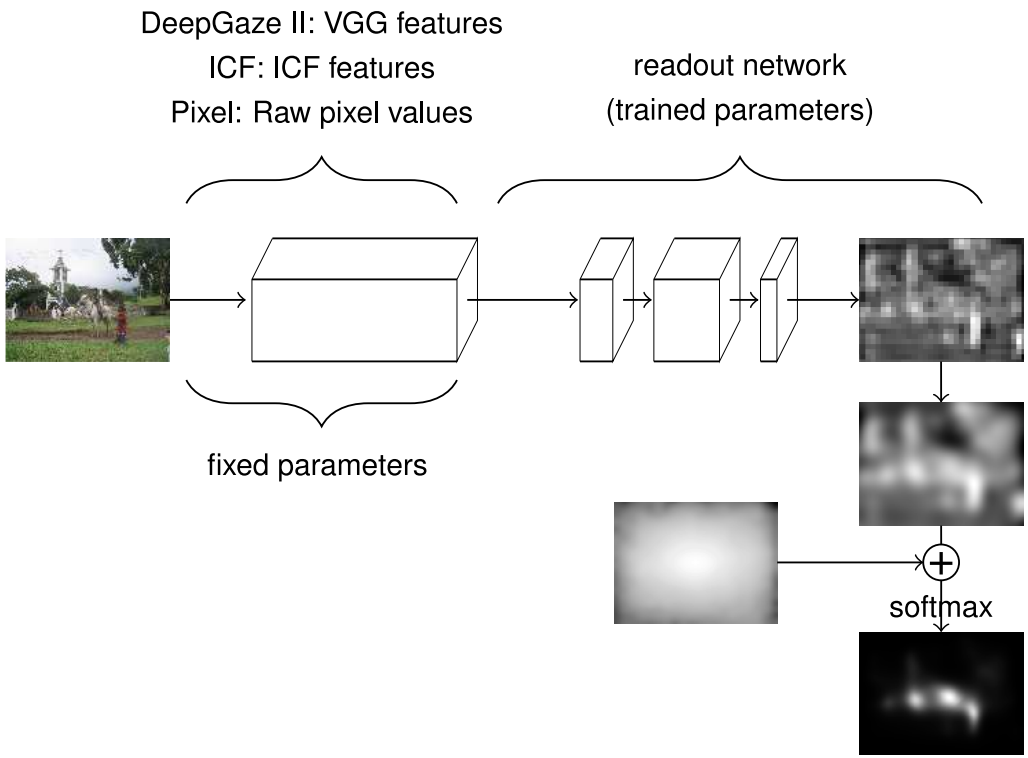
リズム現象 (3)





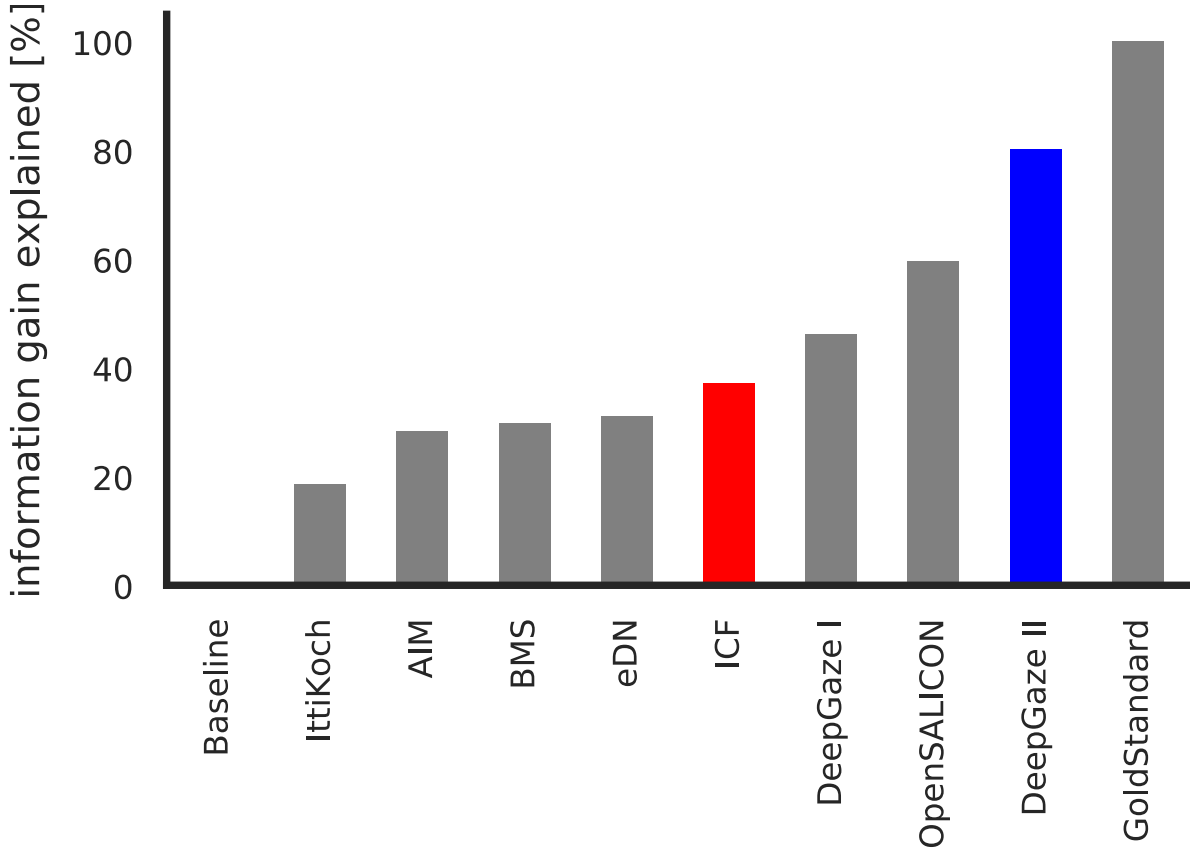
From (Buschman and Kastner 2015) Fig. 3a, Fig. 6}

DeepGaze II



From (Kümmerer et al. 2017) Fig. 2

DeepGaze II (2)



From (Kümmerer et al. 2017) Fig. 2

DeepGazeII より成績の良い最右の棒は人間の眼球運動データ

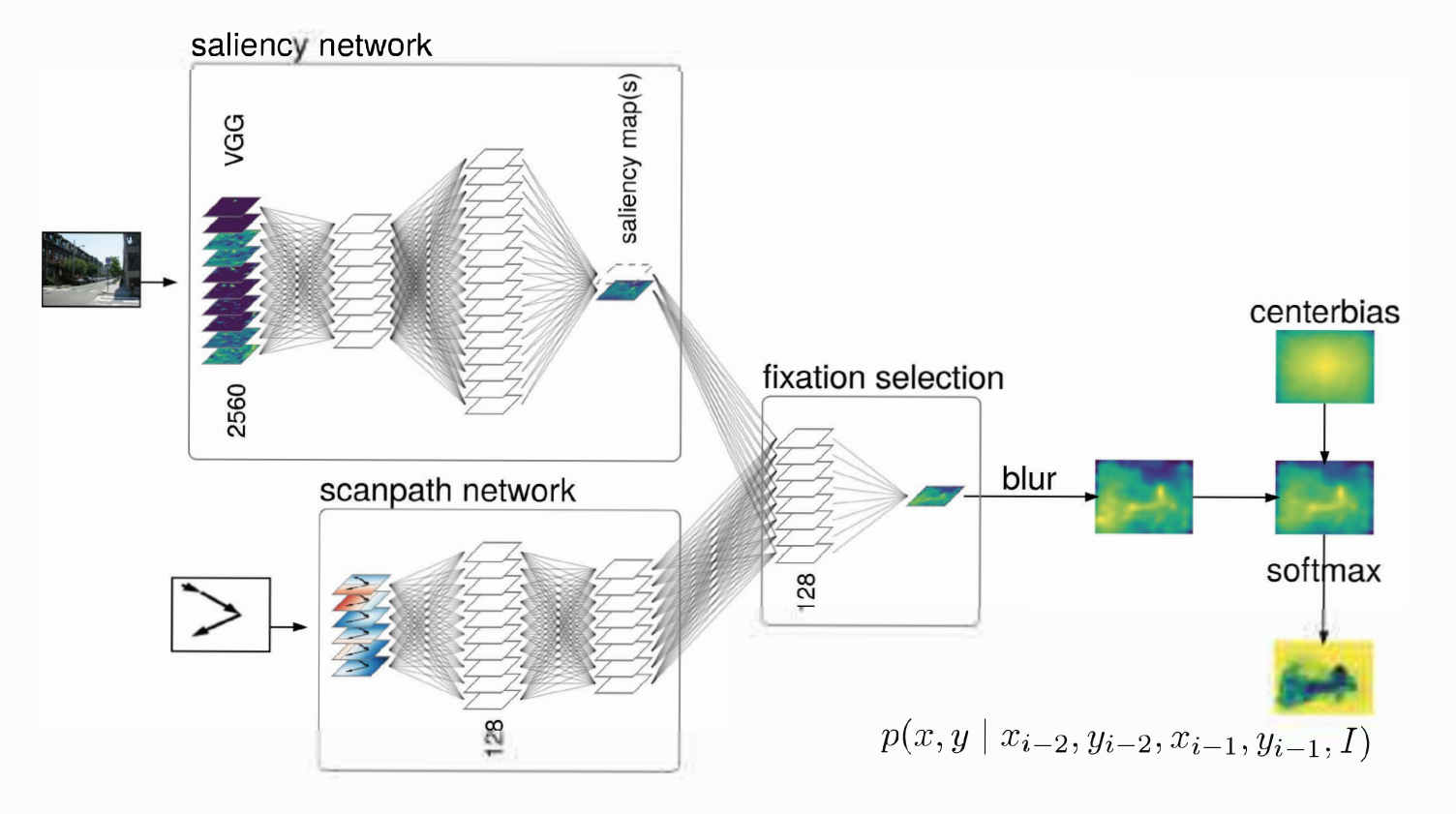
DeepGaze II (3)

Model	IG	IGE	AUC	sAUC	NSS
Centerbias	0.00	0.0	79.6	50.0	1.22
Pixel	0.13	10.7	81.2	60.2	1.38
IttiKoch [16]	0.23	18.6	82.3	64.1	1.41
AIM [6]	0.27	22.6	82.9	65.6	1.50
eDN [48]	0.38	31.1	83.8	68.7	1.61
ICF	0.45	37.2	84.4	70.1	1.74
DeepGaze I [32]	0.56	46.1	85.8	73.0	1.92
OpenSALICON [46]	0.73	59.7	86.4	74.2	2.14
DeepGaze II	**0.98**	**80.3**	**88.3**	**77.7**	**2.48**
Gold Standard	1.22	100.0	89.9	81.2	2.82

From (Kümmerer et al. 2017) Fig. 3

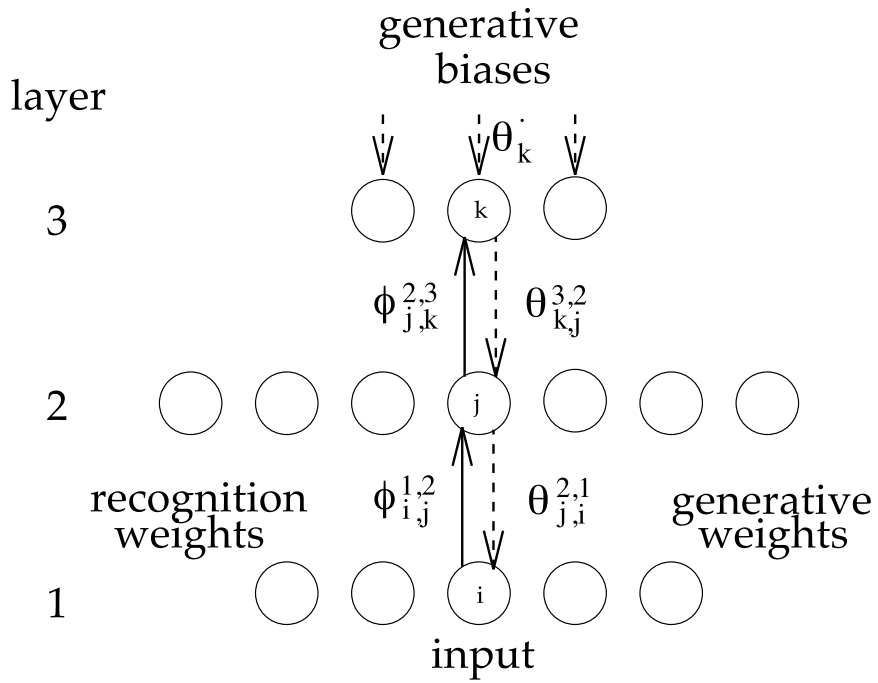
IG: 情報ゲイン, IGE: 修正情報ゲイン, ACU: area under the ROC curve, sAUC: シャッフル精度, NSS: 正規化済キャンパス顕在性 normalized scanpath saliency

DeepGaze III



From (Kümmerer, Wallis, and Bethge 2019) Fig. 1

ヘルムホルツマシン



(Dayan et al. 1995);(Hinton et al. 1995)

ヘルムホルツマシン

$$\log p(d|\theta) = -\sum Q_a E_a - \sum Q_a \log Q_a + \sum Q_a \log\left(\frac{Q_a}{P_a}\right) \tag{8}$$

$$= -F(d;\theta,Q) + \sum_a Q_a \log\left(\frac{Q_a}{P_a}\right)$$

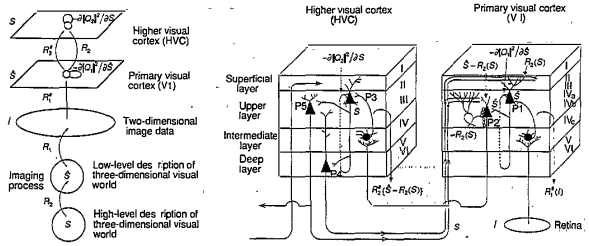
$$q^{(l)}\left(\phi, \mathbf{s}^{(l-1)}\right) = \sigma\left(\sum s^{l-1}\phi^{(l-1,l)}\right) \tag{9}$$

$$Q_\alpha(\phi, d) = \prod\prod \left[q^{(l)}\left(\phi, \mathbf{s}^{(l-1)}\right)\right]^{s^l} \left[1 - q^{(l)}\left(\phi, \mathbf{s}^{(l-1)}\right)\right]^{1-s} \tag{10}$$

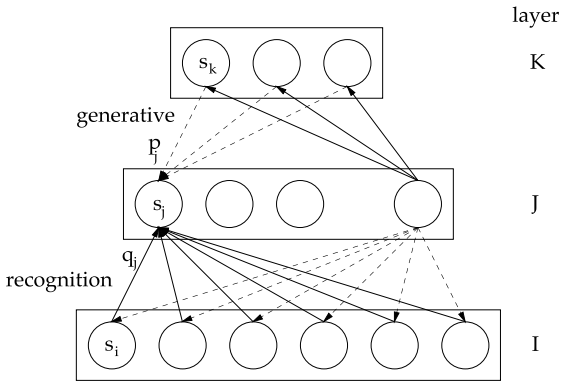
$$p_j^{(l)}\left(\theta, \mathbf{s}^{(l+1)}\right) = \sigma\left(\sum s^{(l+1)}\theta^{(l+1)}\right) \tag{11}$$

$$p(\alpha|\theta) = \prod\prod \left[p_j^{(l)}\left(\theta, \mathbf{s}^{(l+1)}\right)\right] \tag{12}$$

モデル: ヘルムホルツマシン



From (Kawato, Hayakawa, and Inui 1993) Fig. 1 より



From (Hinton et al. 1995) Fig. 1 より

- 上位層は下位層からの情報をサンプリング → 認識形成
- 下位層は上位層からの情報を受けとる → 情報再構成

ボトムアップ処理による認識とトップダウン処理による(こう見えるはずだという思い込みの)生成を n 回繰り返す → $n = 1, 2, \ldots$

定式化

思い込みの印象 α と入力画像 d を用いて%%の記述長は，単なる前隠れ層ユニットの記述損失であり

$$C(\alpha, d) = C(\alpha) + C(d|\alpha) \tag{13}$$
$$= \sum_{\ell \in L} \sum_{j \in \ell} C(s_j^\alpha) + \sum_i C(s_i^d|\alpha)$$

上式を用いて結合係数の更新を行う

$$\Delta w_{kj} = \epsilon s_k^\alpha \left(s_j^\alpha - p_j^\alpha \right), \tag{14}$$

$$C(d) = \sum_\alpha Q(\alpha|d) C(\alpha, d) - \left[- \sum_\alpha Q(\alpha|d) \log Q(\alpha|d) \right]. \tag{15}$$

$$p(\alpha|d) = \frac{e^{-C(\alpha,d)}}{\sum_\beta e^{-C(\beta,d)}} \tag{16}$$

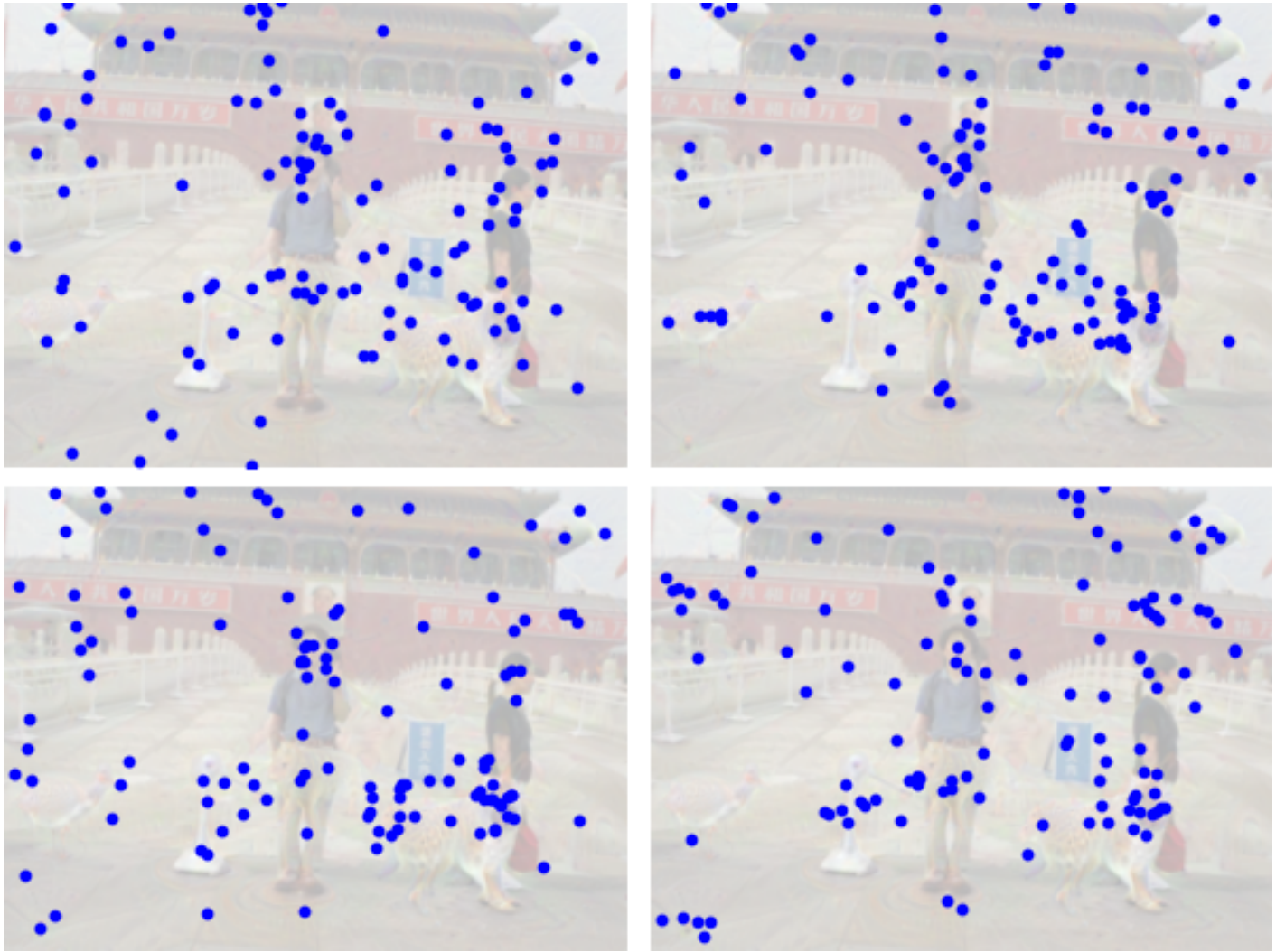
$$\Delta s_{j,t+1} = \epsilon s_{j,t}^\gamma (s_{j,t}^\gamma - q_{j,t}^\gamma) \tag{17}$$

全体の良い表象が得られるまで，すなわち下位層の活性を再構築するように複数回繰り返す

計算例

計算例

計算例 (2) 眼球運動のサンプリング}



文献

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. “Neural Machine Translation by Jointly Learning to Align and Translate.” In *Proceedings in the International Conference on Learning Representations (ICLR)*, edited by Yoshua Bengio and Yann LeCun. San Diego, CA, USA.

Bichot, Narcisse P., Matthew T. Heard, Ellen M. DeGennaro, and Robert Desimone. 2015. “A Source for Feature-Based Attention in the Prefrontal Cortex.” *Neuron* 88 (November): 832–44.

Bloom, Floyd E., and Arlyne Lazerson. 1988. *Brain, Mind, and Behavior*. 2nd ed. New York, NY: Freeman.

Borji, Ali, and Laurent Itti. 2013. “State-of-the-Art in Visual Attention Modeling.” *IEEE Transaction on Pattern Analysis and Machine Intelligence* 35 (1): 185–207.

Broadbent, Donald E. 1958. *Perception and Communication*. Oxford,UK: Pergamon.

Buschman, Timothy J., and Sabine Kastner. 2015. “From Behavior to Neural Dynamics: An Integrated Theory of Attention.” *Neuron* 88 (October): 127–44.

Chen, Yunpeng, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. 2018. “A^2-Nets: Double Attention Networks.” In *Advances in Neural Information Processing Systems 31*, edited by S Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, 352–61. Curran Associates, Inc. http://papers.nips.cc/paper/7318-a2-nets-double-attention-networks.pdf.

Cordonnier, Jean-Baptiste, Andreas Loukas, and Martin Jaggi. 2020. “ON the Relationship Between Self-Attention and Convolutional Layers.” *ArXiv Preprint* [cs.LG] (1911.035842). https://arxiv.org/1911.03584/.

Crick, Francis. 1984. “Function of the Thalamic Reticular Complex: The Search Light Hypothesis.” *Proceedings of the National Academy of Sciences* 81 (July): 4586–90.

Dayan, Peter, Geoffrey E. Hinton, Radford M. Neal, and Richard S. Zemel. 1995. “The Helmholtz Machine.” *Neural Computation* 7: 889–904.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” *arXiv Preprint*.

Duncan, John, and Glyn W. Humphreys. 1989. “Visual Search and Stimulus Similarity.” *Psychological Review* 96 (3): 433–58.

Elman, Jeffrey L. 1990. “Finding Structure in Time.” *Cognitive Science* 14: 179–211.

Eriksen, Charles W., and James D. St.James. 1986. “Visual Attention Within and Around the Field of Focal Attention: A Zoom Lens Model.” *Perception and Psychophysics* 40 (4): 225–40.

Fiebelkorn, Ian C., Yuri B. Saalmann, and Sabine Kastner. 2013. “Rhythmic Sampling Within and Between Objects Despite Sustained Attention at a Cued Location.” *Current Biology* 23 (December): 2553–8.

Friston, Karl J, Klaas Enno Stephan, Read Montague, and Raymond J Dolan. 2014. “Computational Psychiatry: The Brain as a Phantastic Organ.” *The Lancet Psychiatry* 1: 148–58.

Gers, Fleix A., Jürgen Schmidhuber, and Fred Cummins. 1999. “Learning to Forget: Continual Prediction with LSTM.” In *Artificial Neural Networks ICANN 99. Ninth International Conference on*, 2:850–55. Edinburgh, Scotland.

Graves, Alex, Greg Wayne, and Ivo Danihelka. 2014. “Neural Turing Machines.” *ArXiv:1410.5401*.

Graves, Alex, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwi nska, Sergio Gómez Colmenarejo, et al. 2016. “Hybrid Computing Using a Neural Network

with Dynamic External Memory." *Nature* 538: 471–76. https://doi.org/10.1038/nature20101.

Greff, Klaus, Rupesh Kumar Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. 2015. "LSTM: A Search Space Odyssey." *ArXiv:1503.04069*.

Heilman, Kennerh M., and Edward Valenstein. 1979. "Mechanisms Underlying Hemispatial Neglect." *The Annals of Neurology* 5 (2): 166–70.

Hewitt, John, and Christopher D. Manning. 2019. "A Structural Probe for Finding Syntax in Word Representations." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4129–38. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1419.

Hinton, Geoffrey E., Peter Dayan, Brendan J. Frey, and Radford M. Neal. 1995. "The "Wake-Sleep" Algorithm for Unsupervised Neural Networks." *Science* 268 (5214): 1158–61.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9: 1735–80.

Hopfield, John Joseph. 1982. "Neural Networks and Physical Systems with Emergent Collective Computational Abilities." *Proceedings of the National Academy of Sciences* 79: 2554–8.

Itti, Laurent, and Ali Borji. 2014. "Computational Models: Bottom-up and Top-down Aspects." In *The Oxford Handbook of Attention*, edited by Anna C. Nobre and Sabine Kastner, 1122–58. Oxford University Press.

———. 2015. "Computational Models of Attention." *ArXiv Preprint*.

Itti, Laurent, and Christof Koch. 2001. "Computational Modelling of Visual Attention." *Nature Reviews Neuroscience* 2: 1–11.

Itti, Laurent, Christof Koch, and Ernst Niebur. 1998. “A Model of Saliency-Based Visual Attention for Rapid Scene Analysis.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (11): 1254–9.

Jones, Karen Spärck. 1972. “A Statistical Interpretation of Term Specificity and Its Application in Retrieval.” *Journal of Documentation* 28 (1): 11–21.

Kawato, Mitsuo, Hideki Hayakawa, and Toshio Inui. 1993. “A Forward-Inverse Optics Model of Reciprocal Connections Between Visual Cortical Areas.” *Network: Computation in Neural Systems* 4 (4): 415–22.

Kim, Junho, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. 2019. “U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation.” *ArXiv Preprint* [cs.CV] (1907.10830).

Kimura, Akisato, Ryo Yonetani, and Takatsugu Hirayama. 2013. “Computational Models of Human Visual Attention and Their Implementations: A Survey.” *IEICE Transactions of Information & Systems* E96-D (3): 562–78.

Knudsen, Eric I. 2007. “Fundamental Components of Attention.” *Annual Revivew of Neuroscience* 30: 57–78.

Koch, Christoh, and Simon Ullman. 1985. “Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry.” *Human Neurobiology* 4: 219–27.

Krauzlis, Richard J., Lee P. Lovejoy, and Alexandre Zénon. 2013. “Superior Colliculus and Visual Spatial Attention.” *Annual Review of Neuroscience* 36 (165–182).

Kümmerer, Matthias, Thomas S.A. Wallis, and Matthias Bethge. 2019. “DeepGaze III: Using Deep Learning to Probe Interactions Between Scene Content and Scanpath History in Fixation Selection.”

In *Proceedings of Cognitive Computational Neuroscience*, 542–45. Berlin, Germany. https://doi.org/https://doi.org/10.32470/CCN.2019.1235-0.

Kümmerer, Matthias, Thomas S. A. Wallis, Leon A. Gatys, and Matthias Bethge. 2017. “Understanding Low- and High-Level Contributions to Fixation Prediction.” In *The IEEE International Conference on Computer Vision (ICCV)*, 4789–98. Venice, Italy.

Lample, Guillaume, and Alexis Conneau. 2019. “Cross-Lingual Language Model Pretraining.” *ArXiv Preprint* 1901.07291v1 [cs.CL].

Liu, Xiaodong, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. “Multi-Task Deep Neural Networks for Natural Language Understanding.” In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4487–96. Florence, Italy: Association for Computational Linguistics.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. “RoBERTa: A Robustly Optimized Bert Pretraining Approach.” *ArXiv Preprint*.

Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. 2015. “Effective Approaches to Attention-Based Neural Machine Translation.” *ArXiv Preprint* cs.CL: 1508.04025.

Manning, Christopher D., and Hinrich Schuütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT press.

Mikolov, Tomaś, Stefan Kombrink, Lukaś Burget, Jan “Honza" Černocký, and Sanjeev Khudanpur. 2011. “Extensions of Recurrent Neural Network Language Model.” In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Prague, Czech Republic.

Mikolov, Tomáš, Martin Karafiát, Lukáš Burget, Jan “Honza” Černocký, and Sanjeev Khudanpur. 2010. “Recurrent Neural Network Based Language Model.” In *Proceedings of INTERSPEECH2010*,

edited by Takao Kobayashi, Keiichi Hirose, and Satoshi Nakamura, 1045–8. Makuhari, JAPAN.

Milanese, Ruggero, Harry Wechsler, Sylvia Gill, Jean-Marc Bost, and Thierry Pun. 1994. “Integration of Bottom-up Integration of Bottom-up and Top-down Cues for Visual Attention Using Non-Linear Relaxation.” In *The Proceedings of CVPR, IEEE – Institute of Electrical and Electronics Engineers*, 781–85. Dallas Texas, USA: Computer Vision; Pattern Recognition (CVPR).

Miller, Earl K., and Jonathan D. Cohen. 2001. “An Integrative Theory of Prefrontal Cortex Function.” *Annual Revivew of Neuroscience* 24 (167–202).

Mishra, Nikhil, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. 2018. “A Simple Neural Attentive Meta-Learner.” *ArXiv Preprint* [cs.AI] (1707.03141).

Mnih, Volodymyr, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. “Recurrent Models of Visual Attention.” In *Advances in Neural Information Processing Systems 27*, edited by Zoubin Ghahramani, Max Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, 2204–12. Curran Associates, Inc. http://papers.nips.cc/paper/5542-recurrent-models-of-visual-attention.pdf.

Monosov, Ilya E., and Kirk G. Thompson. 2009. “Frontal Eye Field Activity Enhances Object Identification During Covert Visual Search.” *Journal of Neurophysiology* 102 (October): 3656–72.

Moore, Tirin, and Marc Zirnsak. 2017. “Neural Mechanisms of Selective Visual Attention.” *Annual Review of Psychology* 68 (January): 47–72. https://doi.org/10.1146/annurev-psych-122414-033400.

Olshausen, Bruno A., Charles H. Anderson, and David C. Van Essen. 1993. “A Neurobiological Model of Visual Attention and Invariant Pattern Recognition Based on Dynamic Routing of Information.” *The Journal of Neuroscience* 13 (11): 4700–4719.

Petersen, Steven E., and Michael I. Posner. 2012. “The Attention System of the Human Brain: 20 Years After.” *Annual Review of Neuroscience* 35: 73–89.

Posner, Michael I., and Steven E. Petersen. 1990. “The Attention System of the Human Brain.” *Annual Review of Neuroscience* 13: 25–42.

Posner, Michel I. 1980. “Orienting of Attention.” *Quarterly Journal of Experimental Psychology* 32: 3–25.

Ramachandran, Prajit, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. 2019. “Stand-Alone Self-Attention in Vision Models.” *ArXiv Preprint* [cs.CV] (1906.05909). https://arxiv.org/1906.059009/.

Salinas, Emilio, and L. F. Abbott. 1997. “Invariant Visual Responses from Attentional Gain Fields.” *Journal of Neurophsiology* 77: 3267–72. https://doi.org/10.1152/jn.1997.77.6.3267.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. “DistilBERT, a Distilled Version of Bert: Smaller, Faster, Cheaper and Lighter.” *ArXiv Preprint*. https://arXiv.org/1910.01108.

Sperry, Roger W. 1961. “Cerebral Organization and Behavior.” *Science* 133: 1749–57.

Sperry, Roger W. 1968. “Hemisphere Deconnection and Unity in Conscious Awareness.” *American Psychologist* 28: 723–33.

Summerfield, Jennifer J., Jöran Lepsien, Darren R. Gitelman, M. Marsel Mesulam, and Anna C. Nobre. 2006. “Orienting Attention Based on Long-Term Memory Experience.” *Neuron* 49: 905–16. https://doi.org/10.1016/j.neuron.2006.01.021.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. “Sequence to Sequence Learning with Neural Networks.” In *Advances in Neural Information Processing Systems (NIPS)*, edited by Zoubin Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, 27:3104–12. Montreal, BC, Canada.

Treisman, Ann. 1964. “Selective Attention in Man.” *British Medical Bulletin* 20: 12–16.

———. 1988. “Feature and Objects: The Fourteenth Bartlett Memorial Lecture.” *The Quarterly Journal of Experimental Psychology* 40A: 201–37.

Treisman, Anne M. 1969. “Strategies and Models of Selective Attention.” *Psychological Review* 76 (3): 282–99.

Treisman, Ann, and George Gelade. 1980. “A Feature Integration Theory of Attention.” *Cognitive Psychology* 12: 97–136.

Treisman, Ann, and J. Souther. 1985. “Search Asymmetry: A Diagnostic for Preattentive Processing of Separable Features.” *Journal of Experimental Psychology: General* 114 (3): 285–310.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, and Lukasz Kaiser. 2017. “Attention Is All You Need.” *arXiv Preprint* [cs.CL] (1706.03762).

Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. “Show and Tell: A Neural Image Caption Generator.” In *Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA.

Wang, Fei, Mengqing Jiang, Chen Qian, Shuo Yang, and Cheng Li. 2017. “Residual Attention Network for Image Classification.” In *Proceedings of International Conference of Computer Vision (ICCV), IEEE International Conference*.

Wang, Wenguan, and Jianbin Shen. 2018. “Deep Visual Attention Prediction.” *IEEE Transactions on Image Processing* 27 (5): 2368–78.

Wang, Xiaolong, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. “Non-Local Neural Networks.” *ArXiv Preprint* [cs.CV]. https://arxiv.org/1711.07971.

Wardak, Claire, Etienne Olivier, and Jean-René Duhamel. 2004. “A Deficit in Covert Attention After Parietal Cortex Inactivation in the Monkey.” *Neuron* 42 (May): 501–8.

Wolfe, Jeremy M. 1994. “Guided Search 2.0 a Revised Model of Visual Search.” *Psychonomic Bulletin and Review* 1 (2): 202–38.

Xu, Kelvin, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.” *ArXiv:1502.03044*.

Zhang, Han, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2019. “Self-Attention Generative Adversarial Networks.” *ArXiv Preprint* [stat.ML] (1805.08318).