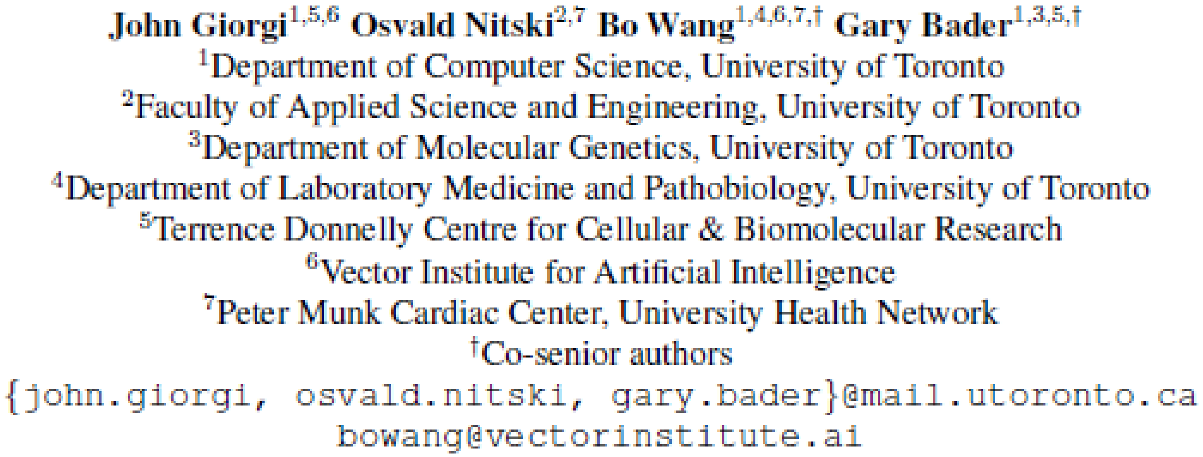
最先端NLP2021

DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations

(ACL 2021)



John Giorgi[1,5,6] Osvald Nitski[2,7] Bo Wang[1,4,6,7,†] Gary Bader[1,3,5,†]
[1]Department of Computer Science, University of Toronto
[2]Faculty of Applied Science and Engineering, University of Toronto
[3]Department of Molecular Genetics, University of Toronto
[4]Department of Laboratory Medicine and Pathobiology, University of Toronto
[5]Terrence Donnelly Centre for Cellular & Biomolecular Research
[6]Vector Institute for Artificial Intelligence
[7]Peter Munk Cardiac Center, University Health Network
[†]Co-senior authors
{john.giorgi, osvald.nitski, gary.bader}@mail.utoronto.ca
bowang@vectorinstitute.ai

発表者：豊田工業大学D3　辻村有輝

特に断りのない限りスライド中の図表は元論文からの引用です

まとめ

- 教師なしで高品質なsentence embeddingを得る手法を提案
  - Quick-Thoughtモデルにかなり近い。訓練事例の作り方が最大の違い
- Siamese network型の学習手法
  - 同一のエンコーダーで各事例をそれぞれエンコードし対照学習
    - 2つの表現ベクトルが同じ”セグメント”由来であればポジティブ
    - 別セグメントや別ドキュメント由来ならネガティブ
- 評価はSentEvalで実施
  - モデルサイズ対しにスケール（base>small）
  - 学習事例数に対してもスケール（ただしMLMの追加学習も必須）

背景

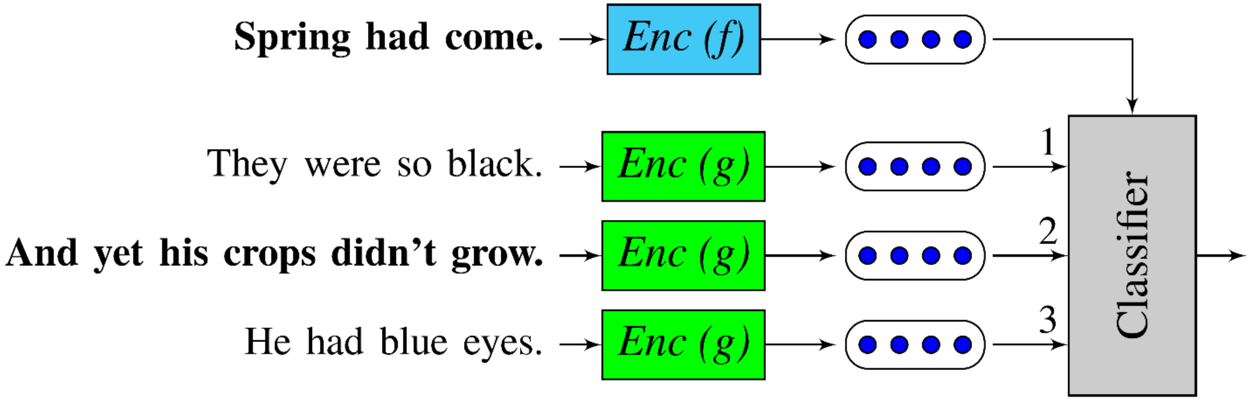
- 高品質な文表現が得られれば様々なタスクに応用可能
- しかしながら現在提案されている高品質な文表現の学習手法は教師あり学習
  - InferSent, Universal Sentence Encoder, Sentence Transformers
  - 大量のラベルありデータを用意するのにコストがかかり、低リソース言語での利用が困難

教師なし学習で高品質な文表現を学習できるようにしたい！
➡近い位置の文かどうかを判定させる教師なし学習を行うことで
高品質な文表現を学習する手法を提案

先行研究｜Quick-Thought

- 提案手法はQuick-Thoughtモデルにかなり近い
  - Quick-Thought：**前後の隣接文**を学習に使用
  - 提案手法：**隣接/内包/一部重複**するようなスパン（必ずしも文ではない）

- Qucik-Thought：教師なしの文表現学習
  - 与えられた文群のうち、基準文の**前後にある文**を**分類**する
  - エンコーダーはシンプルなGRUで最後のセル出力を表現ベクトルととる



$$p(s_{\mathrm{cand}}|s, S_{\mathrm{cand}}) = \frac{\exp[c(f(s), g(s_{\mathrm{cand}}))]}{\sum_{s' \in S_{\mathrm{cand}}} \exp[c(f(s), g(s'))]}$$

引用元：Logeswaran and Lee. 2018. An efficient framework for learning sentence representations. ICLR 2018.

先行研究｜Skip-Thought

- Skip-Thought：教師なしの文表現学習
  - Quick-Thoughtの先行研究
  - 文表現から**前後の隣接文**の**生成**を学習させる

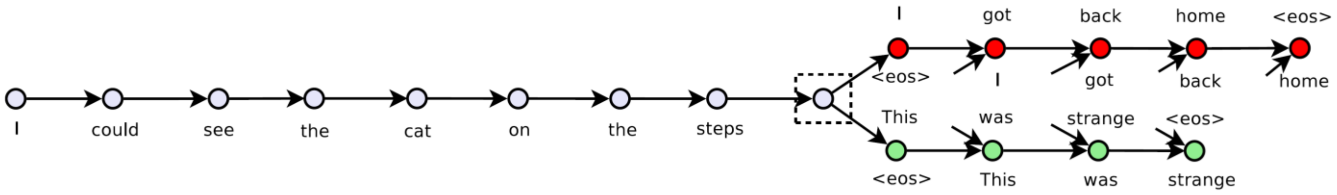
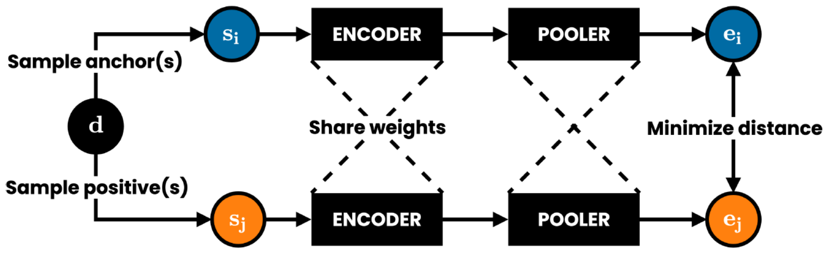


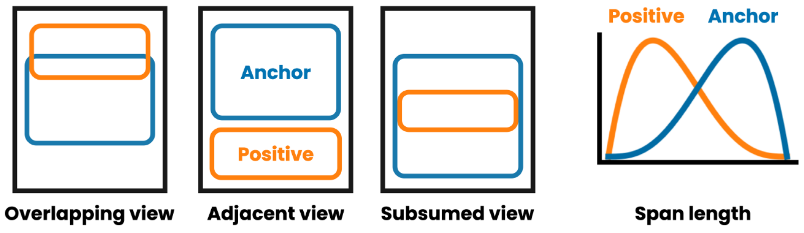
Figure 1: The skip-thoughts model. Given a tuple (s_{i-1}, s_i, s_{i+1}) of contiguous sentences, with s_i the i-th sentence of a book, the sentence s_i is encoded and tries to reconstruct the previous sentence s_{i-1} and next sentence s_{i+1}. In this example, the input is the sentence triplet *I got back home. I could see the cat on the steps. This was strange.* Unattached arrows are connected to the encoder output. Colors indicate which components share parameters. $\langle$eos$\rangle$ is the end of sentence token.

引用元：Ryan Kiros et al. Skip-Thought Vectors. NIPS 2015.

手法｜おおまかな概要

- Siamese network型の対照学習
  - 同一のエンコーダーで各事例をそれぞれエンコードし各表現ベクトルを得る
  - 学習はcontrastive learning（対照学習）で行う
    - 表現ベクトルペアがポジティブなら近付け、ネガティブなら遠ざける
    - 2つの表現ベクトルが同じ”セグメント”由来であればポジティブ
    - 別セグメントや別ドキュメント由来ならネガティブ

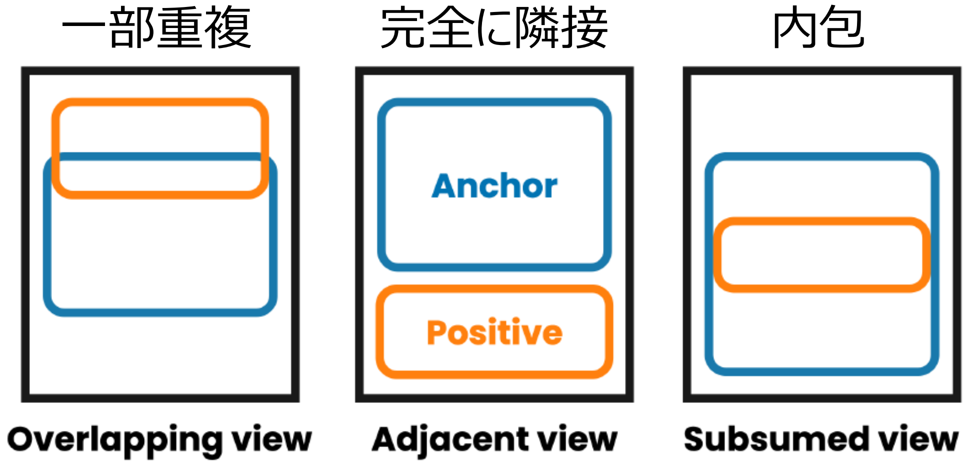




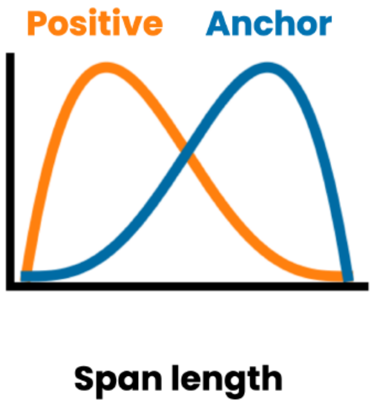
手法｜事例ペア作成

- ドキュメント中の同一セグメント内からサンプリングされていると ポジティブなペア
- 別セグメントや、別のドキュメントからサンプルされていると ネガティブなペア

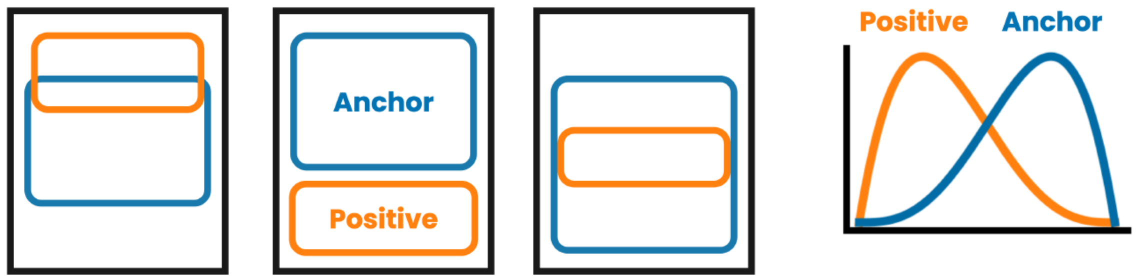
同一セグメントのバリエーションは3種類



各文長は以下のβ分布より決定
正例はAnchorの決定→Positiveの決定の順で作成していく



手法｜事例ペア作成



1. ドキュメントをtokenize

$x^d = (x_1, x_2, \ldots, x_n)$　d：ドキュメントのインデックス　n：トークン数

2. ドキュメント内からランダムにAnchorとなる文を抽出する

$p_{\text{anchor}} \sim \text{Beta}(\alpha = 4, \beta = 2)$ ⬅ 1に近い値が出やすい

$\ell_{\text{anchor}} = \lfloor p_{\text{anchor}} \times (\ell_{\text{max}} - \ell_{\text{min}}) + \ell_{\text{min}} \rfloor$ ⬅ アンカー文長。$[l_{min}, l_{\max}]$の範囲。

$s_i^{\text{start}} \sim \{0, \ldots, n - \ell_{\text{anchor}}\}$ ⬅ 一様分布から始点をサンプリング

$s_i^{\text{end}} = s_i^{\text{start}} + \ell_{\text{anchor}}$

$s_i = x^d_{s_i^{\text{start}}:s_i^{\text{end}}}$

$l_{\min}$：最小文長=32
$l_{\max}$：最大文長=512

3. Anchorと同一セグメントとなる短いPositive文を抽出する

$p_{\text{anchor}} \sim \text{Beta}(\alpha = 2, \beta = 4)$ ⬅ 0に近い値が出やすい

$\ell_{\text{positive}} = \lfloor p_{\text{positive}} \times (\ell_{\text{max}} - \ell_{\text{min}}) + \ell_{\text{min}} \rfloor$

$s_{i+pAN}^{\text{start}} \sim \{s_i^{\text{start}} - \ell_{\text{positive}}, \ldots, s_i^{\text{end}}\}$ ⬅ やはり一様分布から始点をサンプリング。両端が選ばれればadjecentだがほぼoverlapかsubsumになるはず

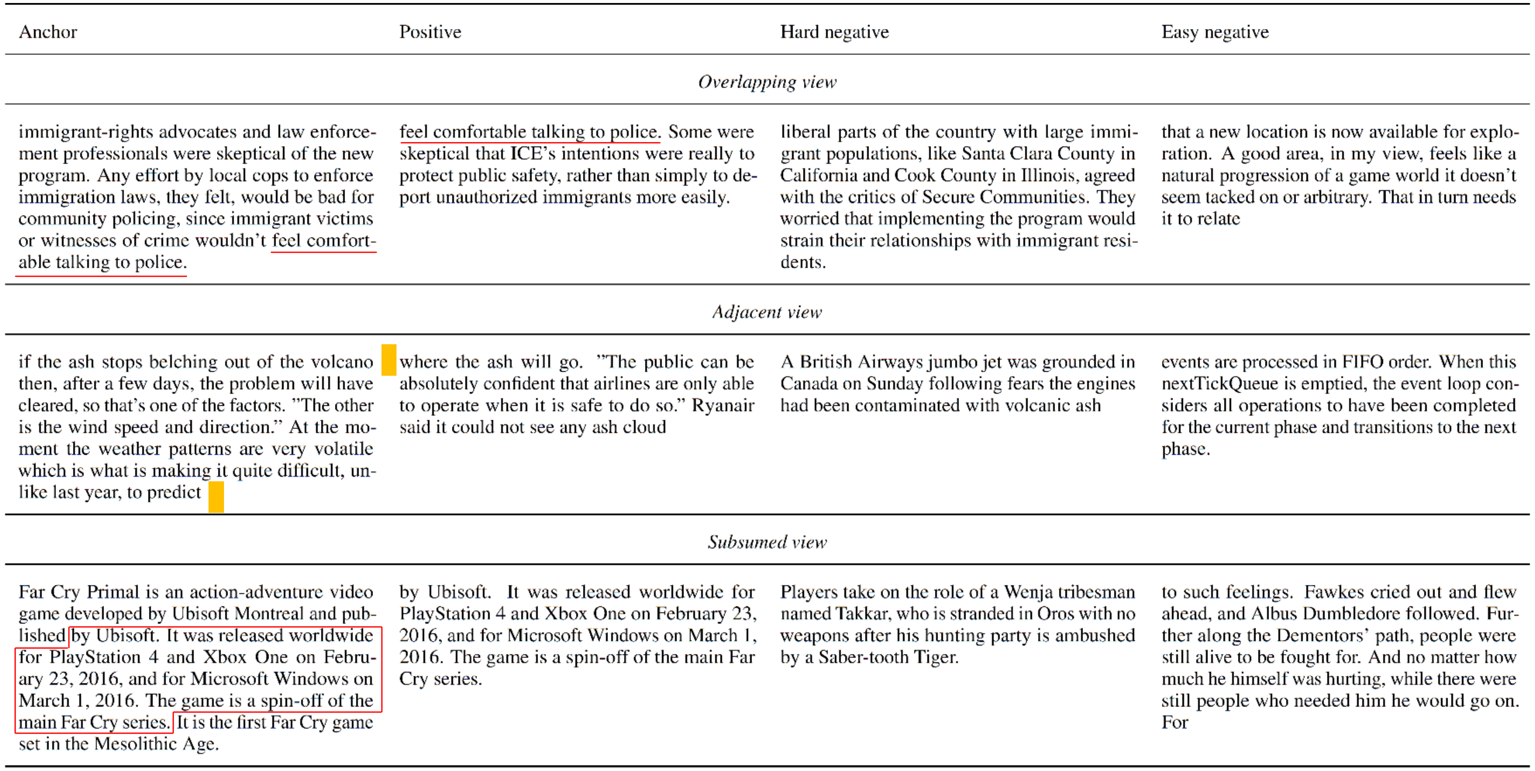
$s_{i+pAN}^{\text{end}} = s_{i+pAN}^{\text{start}} + \ell_{\text{positive}}$

$s_{i+pAN} = x^d_{s_{i+pAN}^{\text{start}}:s_{i+pAN}^{\text{end}}}$

A：ドキュメントごとにA個のAnchorを抽出する。実験では2
P：各アンカーごとにN個のPositiveを抽出する。実験では2
N：ミニバッチサイズ

手法｜事例ペア作成

- 実際の例
  - 文表現の学習のための手法だが学習事例は文ではなくランダムなスパン
  - 同一ドキュメントの別セグメントから抽出されたHard negativeが存在



Anchor	Positive	Hard negative	Easy negative
Overlapping view			
immigrant-rights advocates and law enforcement professionals were skeptical of the new program. Any effort by local cops to enforce immigration laws, they felt, would be bad for community policing, since immigrant victims or witnesses of crime wouldn't feel comfortable talking to police.	feel comfortable talking to police. Some were skeptical that ICE's intentions were really to protect public safety, rather than simply to deport unauthorized immigrants more easily.	liberal parts of the country with large immigrant populations, like Santa Clara County in California and Cook County in Illinois, agreed with the critics of Secure Communities. They worried that implementing the program would strain their relationships with immigrant residents.	that a new location is now available for exploration. A good area, in my view, feels like a natural progression of a game world it doesn't seem tacked on or arbitrary. That in turn needs it to relate
Adjacent view			
if the ash stops belching out of the volcano then, after a few days, the problem will have cleared, so that's one of the factors. "The other is the wind speed and direction." At the moment the weather patterns are very volatile which is what is making it quite difficult, unlike last year, to predict	where the ash will go. "The public can be absolutely confident that airlines are only able to operate when it is safe to do so." Ryanair said it could not see any ash cloud	A British Airways jumbo jet was grounded in Canada on Sunday following fears the engines had been contaminated with volcanic ash	events are processed in FIFO order. When this nextTickQueue is emptied, the event loop considers all operations to have been completed for the current phase and transitions to the next phase.
Subsumed view			
Far Cry Primal is an action-adventure video game developed by Ubisoft Montreal and published by Ubisoft. It was released worldwide for PlayStation 4 and Xbox One on February 23, 2016, and for Microsoft Windows on March 1, 2016. The game is a spin-off of the main Far Cry series. It is the first Far Cry game set in the Mesolithic Age.	by Ubisoft. It was released worldwide for PlayStation 4 and Xbox One on February 23, 2016, and for Microsoft Windows on March 1, 2016. The game is a spin-off of the main Far Cry series.	Players take on the role of a Wenja tribesman named Takkar, who is stranded in Oros with no weapons after his hunting party is ambushed by a Saber-tooth Tiger.	to such feelings. Fawkes cried out and flew ahead, and Albus Dumbledore followed. Further along the Dementors' path, people were still alive to be fought for. And no matter how much he himself was hurting, while there were still people who needed him he would go on. For

手法｜モデル構造・損失関数

- アンカー文の表現ベクトルはBERT→word-wiseなaverage pooling

$$\boldsymbol{e}_i = g\big(f(s_i)\big)$$

f:BERT　g:average pooling
実験ではRoBERTaかDistilRoBERTaを使用

- positive文は各アンカーごとにN個あるのでそれらの平均を一つのpositive表現として使う

$$\boldsymbol{e}_{i+AN} = \frac{1}{P}\sum_{p=1}^{P} g\left(f\left(s_{i+pAN}\right)\right)$$

- 損失は温度付き交差エントロピー

$$l(i,j) = -\log\frac{\exp(\cos(\boldsymbol{e}_i,\boldsymbol{e}_j)/\tau)}{\sum_{k=1,k\neq i}^{2AN}\exp(\cos(\boldsymbol{e}_i,\boldsymbol{e}_k)/\tau)}$$

⬅ i 基準のcos類似度のうち j に対してはどれぐらい大きいか

$$\mathcal{L}_{\text{contrastive}} = \sum_{i=1}^{AN} \ell(i, i+AN) + \ell(i+AN, i)$$

⬅ Anchor (i)基準とPositive ($i+AN$) 基準の合算

- 実際の学習ではMasked Language Modelもfine-tuning： $\mathcal{L} = \mathcal{L}_{\text{contrastive}} + \mathcal{L}_{\text{MLM}}$

実験

- SentEvalで評価
  - 18のDownstream taskと10のProbing taskで構成
  - タスクごとに学習が必要なものと不要なものに分かれている
    - 学習が必要なタスクではエンコードした表現ベクトルと入力とするロジスティック回帰やMLPなどを学習させる

Downstream task	Dataset
Binary/Multi-class classification	CR, MR, MPQA, SUBJ SST2, SST5, TREC
Entailment and semantic relatedness	SNLI, SICK-E, SICK-R, STS-B
Semantic textual similarity	STS12,13,14,15,16
Paraphrase detection	MRPC
Caption-Image retrieval	COCO

Probing tasks
Sentence length (SentLen)
Word content (WC)
Tree depth (TreeDepth)
Bigram Shift (BShift)
Top Constituents (TopConst)

Probing tasks
Tense
Subject number (SubjNum)
Object number (ObjNum)
Semantic odd man out (SOMO)
Coordinate inversion (CoordInv)

実験

- ベースライン
  - 教師あり
    - InferSent
    - Universal Sentence Encoder
    - Sentence Transformers
  - 教師なし
    - Word-wise averaged GloVe/fastText
    - QuickThoughts
    - Transformer (w/o supervised pretraining. i.e., only MLM)

結果

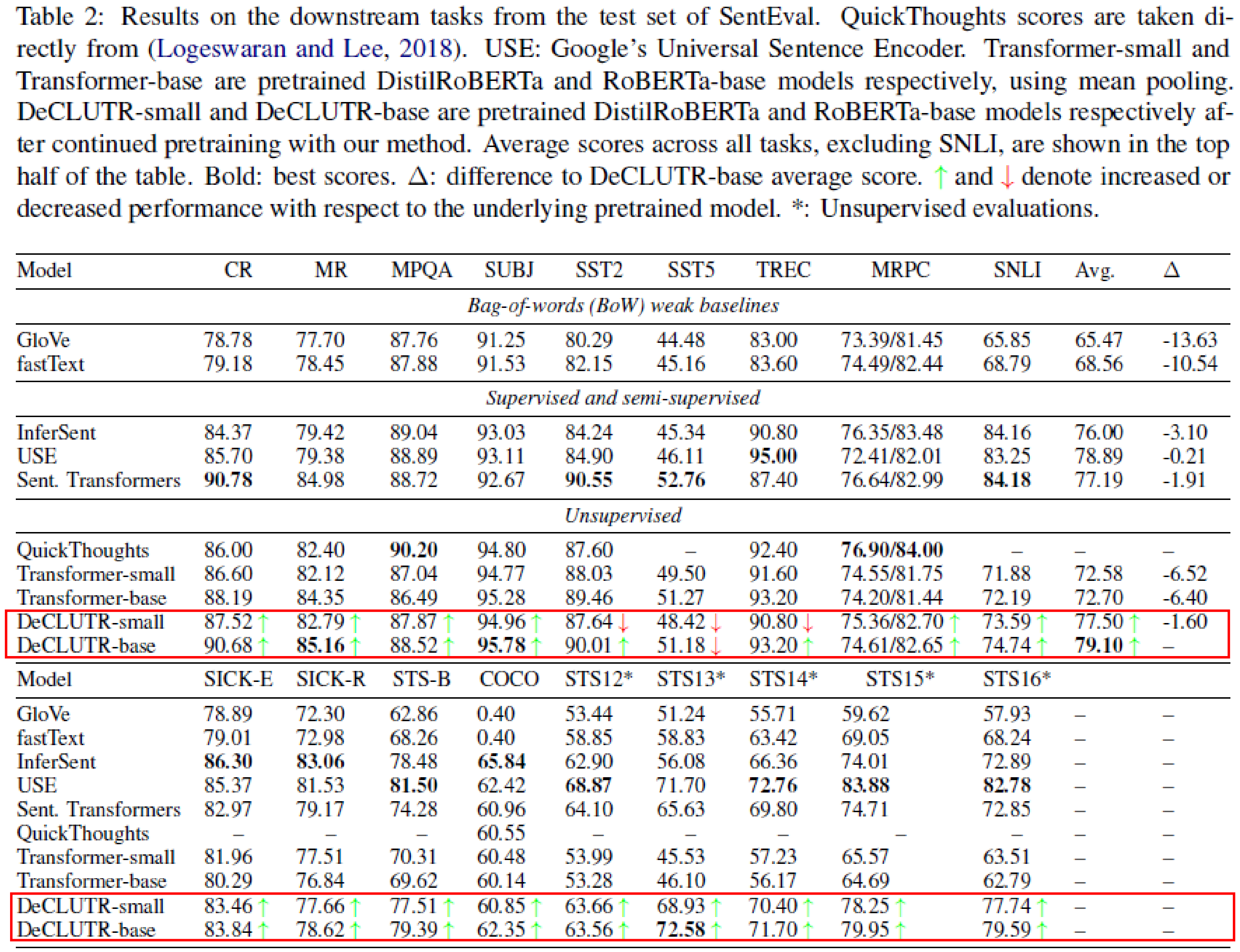


Table 2: Results on the downstream tasks from the test set of SentEval. QuickThoughts scores are taken directly from (Logeswaran and Lee, 2018). USE: Google's Universal Sentence Encoder. Transformer-small and Transformer-base are pretrained DistilRoBERTa and RoBERTa-base models respectively, using mean pooling. DeCLUTR-small and DeCLUTR-base are pretrained DistilRoBERTa and RoBERTa-base models respectively after continued pretraining with our method. Average scores across all tasks, excluding SNLI, are shown in the top half of the table. Bold: best scores. Δ: difference to DeCLUTR-base average score. ↑ and ↓ denote increased or decreased performance with respect to the underlying pretrained model. *: Unsupervised evaluations.

Model	CR	MR	MPQA	SUBJ	SST2	SST5	TREC	MRPC	SNLI	Avg.	Δ
Bag-of-words (BoW) weak baselines											
GloVe	78.78	77.70	87.76	91.25	80.29	44.48	83.00	73.39/81.45	65.85	65.47	-13.63
fastText	79.18	78.45	87.88	91.53	82.15	45.16	83.60	74.49/82.44	68.79	68.56	-10.54
Supervised and semi-supervised											
InferSent	84.37	79.42	89.04	93.03	84.24	45.34	90.80	76.35/83.48	84.16	76.00	-3.10
USE	85.70	79.38	88.89	93.11	84.90	46.11	**95.00**	72.41/82.01	83.25	78.89	-0.21
Sent. Transformers	**90.78**	84.98	88.72	92.67	**90.55**	**52.76**	87.40	76.64/82.99	**84.18**	77.19	-1.91
Unsupervised											
QuickThoughts	86.00	82.40	**90.20**	94.80	87.60	–	92.40	**76.90/84.00**	–	–	–
Transformer-small	86.60	82.12	87.04	94.77	88.03	49.50	91.60	74.55/81.75	71.88	72.58	-6.52
Transformer-base	88.19	84.35	86.49	95.28	89.46	51.27	93.20	74.20/81.44	72.19	72.70	-6.40
DeCLUTR-small	87.52 ↑	82.79 ↑	87.87 ↑	94.96 ↑	87.64 ↓	48.42 ↓	90.80 ↓	75.36/82.70 ↑	73.59 ↑	77.50 ↑	-1.60
DeCLUTR-base	90.68 ↑	**85.16** ↑	88.52 ↑	**95.78** ↑	90.01 ↑	51.18 ↓	93.20 ↑	74.61/82.65 ↑	74.74 ↑	**79.10** ↑	–
Model	SICK-E	SICK-R	STS-B	COCO	STS12*	STS13*	STS14*	STS15*	STS16*		
GloVe	78.89	72.30	62.86	0.40	53.44	51.24	55.71	59.62	57.93	–	–
fastText	79.01	72.98	68.26	0.40	58.85	58.83	63.42	69.05	68.24	–	–
InferSent	**86.30**	**83.06**	78.48	**65.84**	62.90	56.08	66.36	74.01	72.89	–	–
USE	85.37	81.53	**81.50**	62.42	**68.87**	71.70	**72.76**	**83.88**	**82.78**	–	–
Sent. Transformers	82.97	79.17	74.28	60.96	64.10	65.63	69.80	74.71	72.85	–	–
QuickThoughts	–	–	–	60.55	–	–	–	–	–	–	–
Transformer-small	81.96	77.51	70.31	60.48	53.99	45.53	57.23	65.57	63.51	–	–
Transformer-base	80.29	76.84	69.62	60.14	53.28	46.10	56.17	64.69	62.79	–	–
DeCLUTR-small	83.46 ↑	77.66 ↑	77.51 ↑	60.85 ↑	63.66 ↑	68.93 ↑	70.40 ↑	78.25 ↑	77.74 ↑	–	–
DeCLUTR-base	83.84 ↑	78.62 ↑	79.39 ↑	62.35 ↑	63.56 ↑	**72.58** ↑	71.70 ↑	79.95 ↑	79.59 ↑	–	–

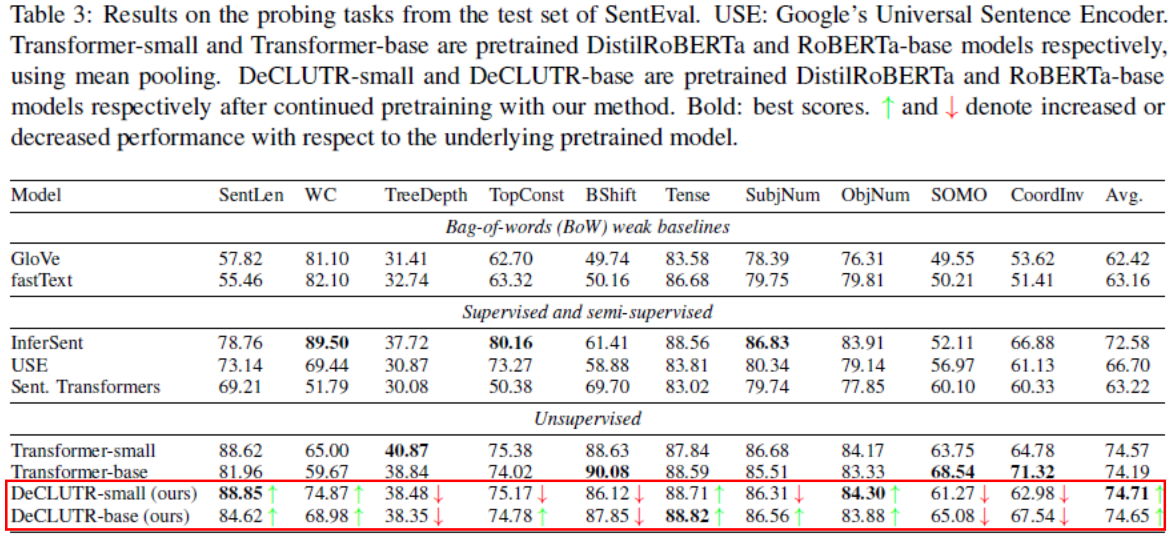


Table 3: Results on the probing tasks from the test set of SentEval. USE: Google's Universal Sentence Encoder. Transformer-small and Transformer-base are pretrained DistilRoBERTa and RoBERTa-base models respectively, using mean pooling. DeCLUTR-small and DeCLUTR-base are pretrained DistilRoBERTa and RoBERTa-base models respectively after continued pretraining with our method. Bold: best scores. ↑ and ↓ denote increased or decreased performance with respect to the underlying pretrained model.

Model	SentLen	WC	TreeDepth	TopConst	BShift	Tense	SubjNum	ObjNum	SOMO	CoordInv	Avg.
Bag-of-words (BoW) weak baselines											
GloVe	57.82	81.10	31.41	62.70	49.74	83.58	78.39	76.31	49.55	53.62	62.42
fastText	55.46	82.10	32.74	63.32	50.16	86.68	79.75	79.81	50.21	51.41	63.16
Supervised and semi-supervised											
InferSent	78.76	**89.50**	37.72	**80.16**	61.41	88.56	**86.83**	83.91	52.11	66.88	72.58
USE	73.14	69.44	30.87	73.27	58.88	83.81	80.34	79.14	56.97	61.13	66.70
Sent. Transformers	69.21	51.79	30.08	50.38	69.70	83.02	79.74	77.85	60.10	60.33	63.22
Unsupervised											
Transformer-small	88.62	65.00	**40.87**	75.38	88.63	87.84	86.68	84.17	63.75	64.78	74.57
Transformer-base	81.96	59.67	38.84	74.02	**90.08**	88.59	85.51	83.33	**68.54**	**71.32**	74.19
DeCLUTR-small (ours)	**88.85** ↑	74.87 ↑	38.48 ↓	75.17 ↓	86.12 ↓	88.71 ↑	86.31 ↓	**84.30** ↑	61.27 ↓	62.98 ↓	**74.71** ↑
DeCLUTR-base (ours)	84.62 ↑	68.98 ↑	38.35 ↓	74.78 ↑	87.85 ↓	**88.82** ↑	86.56 ↑	83.88 ↑	65.08 ↓	67.54 ↓	74.65 ↑

- 教師ありモデルと遜色ない性能
- モデルサイズが大きい方が基本的によい（DeCLUTR-base>small）
  - base：RoBERTa
  - small：DistilRoBERTa

結果

- 各ドキュメントごとに複数セグメントを作成した方がいい
  - i.e. Hard negativeがあった方が良い
- AnchorごとのPositive数はあまり影響がない
- 完全な隣接のみは性能が悪い
  - 完全な隣接しか正例にならない先行研究と対照的

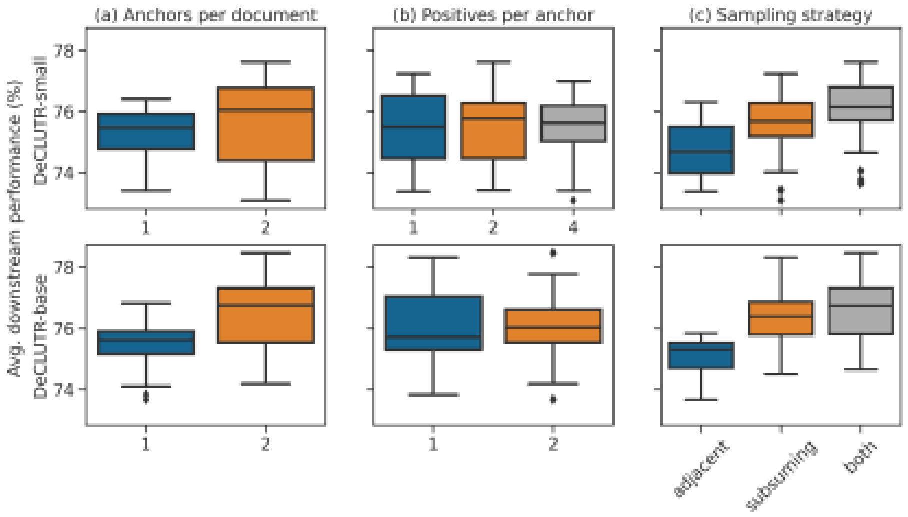


Figure 2: Effect of the number of anchor spans sampled per document (a), the number of positive spans sampled per anchor (b), and the sampling strategy (c). Averaged downstream task scores are reported from the validation set of SentEval. Performance is computed over a grid of hyperparameters and plotted as a distribution. The grid is defined by all permutations of number of anchors $A = \{1, 2\}$, number of positives $P = \{1, 2, 4\}$, temperatures $\tau = \{5 \times 10^{-3}, 1 \times 10^{-2}, 5 \times 10^{-2}\}$ and learning rates $\alpha = \{5 \times 10^{-5}, 1 \times 10^{-4}\}$. $P = 4$ is omitted for DeCLUTR-base as these experiments did not fit into GPU memory.

学習事例の数が増えるほど性能向上

- Contrastive loss単独だと学習事例を増やすとむしろ劣化
- MLM単独だと性能変動はほぼなし

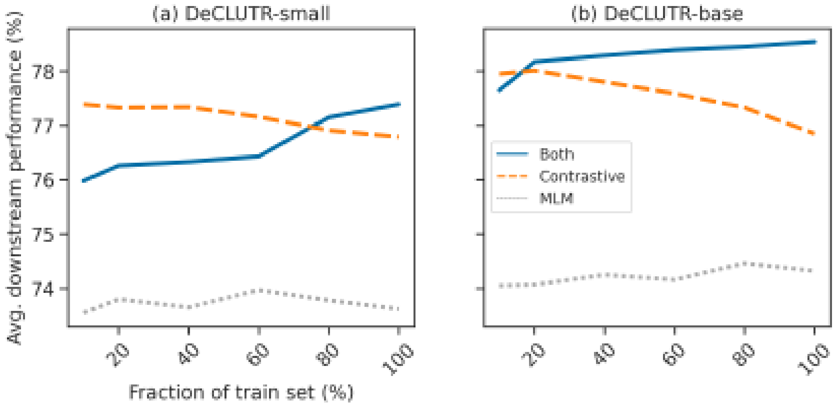


Figure 3: Effect of training objective, train set size and model capacity on SentEval performance. DeCLUTR-small has 6 layers and ∼82M parameters. DeCLUTR-base has 12 layers and ∼125M parameters. Averaged downstream task scores are reported from the validation set of SentEval. 100% corresponds to 1 epoch of training with all 497,868 documents from our OpenWebText subset.

感想

- 正例はほぼoverlapする箇所がある
  - 同一セグメントかどうかはほぼoverlapを見つける問題になるのでは？
  - Overlapを見つける問題になっているとすると、固定長となる文表現ベクトル内に全情報を取りこぼしなく埋め込みたいはずなので、文表現として良い性質を持つかもしれない
  - ごくまれ（2/(512-32)程度）に現れるadjacentはoverlapなしで判定しなければならず、これが適度な訓練難易度の向上に寄与している？

まとめ（再掲）

- 教師なしで高品質なsentence embeddingを得る手法を提案
  - Quick-Thoughtモデルにかなり近い。訓練事例の作り方が最大の違い

- Siamese network型の学習手法
  - 同一のエンコーダーで各事例をそれぞれエンコードし対照学習
    - 2つの表現ベクトルが同じ”セグメント”由来であればポジティブ
    - 別セグメントや別ドキュメント由来ならネガティブ

- 評価はSentEvalで実施
  - モデルサイズ対しにスケール（base>small）
  - 学習事例数に対してもスケール（ただしMLMの追加学習も必須）