第14回最先端NLP勉強会（2022.9.26）

https://aclanthology.org/2022.acl-long.75/

ACL 2022

A Closer Look at How Fine-tuning Changes BERT

Yichu Zhou
School of Computing
University of Utah
flyaway@cs.utah.edu

Vivek Srikumar
School of Computing
University of Utah
svivek@cs.utah.edu

紹介者：笹野遼平（名大）

<このフォーマット>は
紹介者によるコメント

論文の概要

- Fine-tuningによりBERTがどのように変化するか調査した論文
- 得られた主な知見
  1. Fine-tuning introduces a divergence between training and test sets, which is not severe enough to hurt generalization in most cases
  2. Fine-tuning adjusts a representation by grouping points with the same label into a small number of clusters (ideally one)
  3. Fine-tuning pushes the clusters of points representing different labels away from each other, thus introducing large separating regions between labels
  4. Fine-tuning for related tasks can also provide useful signal for the target task by altering the distances between clusters representing different labels
  5. Fine-tuning does not change the higher layers arbitrarily

準備

Probing Methods

- Classifiers as Probes
  - 凍結したembeddingsを利用し分類器を訓練することで、各タスクの情報をどのくらいencodeしているか検証
  - 分類器はFine-tuningとは独立に構成、隠れ層サイズは{32, 64, 128,256}×{32, 64, 128, 256}の2層をGrid Search
- DIRECTPROBE: Probing the Geometric Structure
  - 同じ著者の先行研究[Zhou&Srikumar, NAACL'21]
  - Labeling taskとrepresentationが与えられた場合に以下を満たすクラスタセットを導入
    - 各クラスタは同じラベルの点だけで構成される
    - 異なるクラスタのconvex hull(凸包)がoverlapしない
      (凸包: 与えられた集合を含む最小の凸集合)

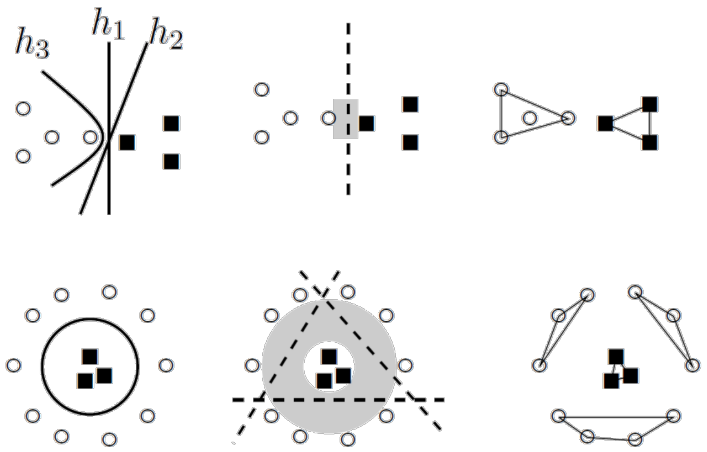


Figure 2: Using the piecewise linear functions to mimic decision boundaries on two different cases. The solid lines on the left are the real decision boundaries for a binary classification problem. The dashed lines in the middle are the piecewise linear functions used to mimic these decision boundaries. The gray area is the region that a separator must cross. The connected points in the right represent the convex regions that the piecewise linear separators lead to.

[Zhou&Srikumar'21]

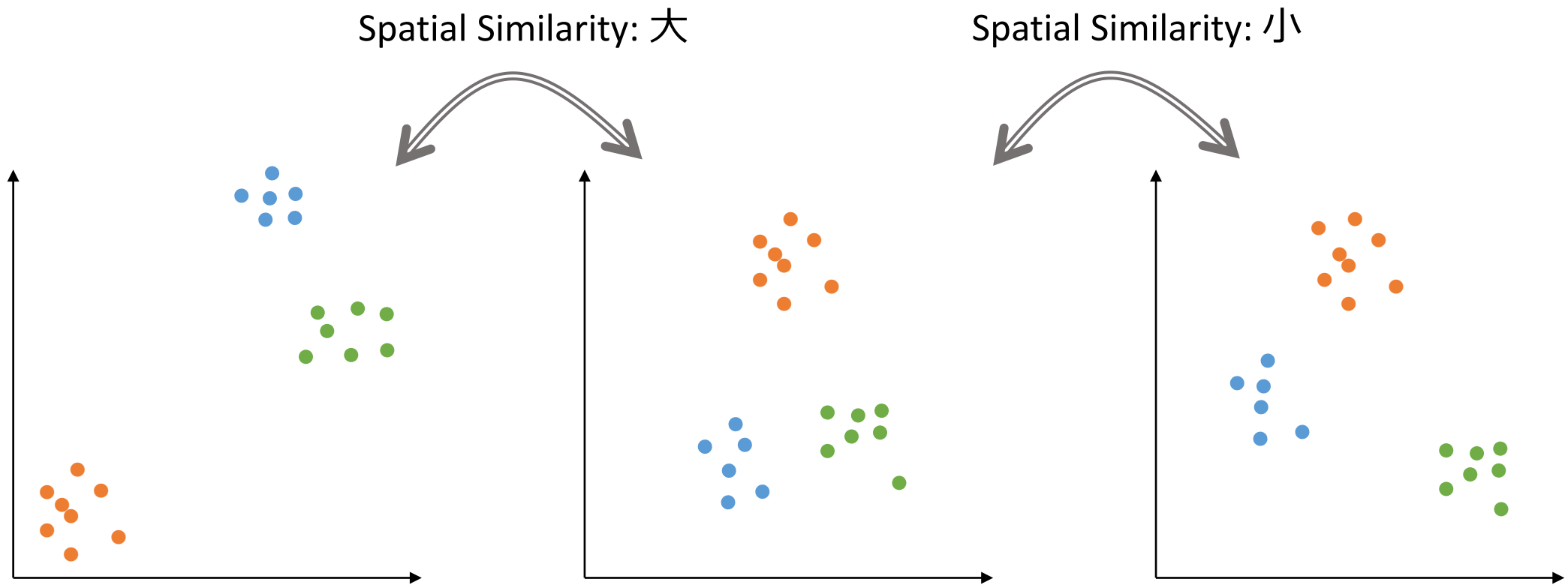
DIRECTPROBE: 着目する3種類の特性

1. Number of Clusters (クラスタの個数)
   - 各タスクに対するrepresentationのlinearityを示す
   - クラスタ数とラベルの種類数が一致する場合は多クラス線形分類器で分類可能
     （linearly separable⇔それ以外の場合は非線形分類器が必要）
2. Distances between Clusters
   - クラスタ間のユークリッド距離
   - max-margin separator (= linear SVM)で算出された marginの2倍がクラスタ間距離(⇐ convex hull)

Representation × ラベル付き
データセットが類似 ⇔
各ラベルの用例で構成される
クラスタ間の距離が類似

3. Spatial Similarity (本論文で新たに導入 <やや分かりにくいので注意>)
   - linearly separableな(クラスタ数＝ラベル数) representation間で定義される空間的類似度
   - representationごとにクラスタ間の距離を列挙したベクトル($|v| = \frac{n(n-1)}{2}$)を作成し、それらのピアソンの相関係数により算出
   - 同一のrepresentationに対する異なるラベル付きデータセットに対しても適用可

Spatial Similarityのイメージ



- 各ラベルに属する点の絶対的な位置の類似性ではなく、各クラスタ間の距離の相対的な類似性に基づきSimilarityを定義

実験で用いるrepresentations

- レイヤー数、次元数等を変えたBERT５種
  - 最大の設定は$\text{BERT}_{\text{base}}$
  - $\text{BERT}_{\text{large}}$はfine-tuning後の状態がunstableだったので含めず

	Layers	#heads	Dim	#Param
$\text{BERT}_{\text{tiny}}$	2	2	128	4.4M
$\text{BERT}_{\text{mini}}$	4	4	256	11.3M
$\text{BERT}_{\text{small}}$	4	8	512	29.1M
$\text{BERT}_{\text{medium}}$	8	8	512	41.7M
$\text{BERT}_{\text{base}}$	12	12	768	110.1M

実験に用いるTasks５つ

1. Part-of-speech tagging (POS): UD*で定義された17種の品詞予測
2. Dependency relation (DEP): UD*で定義された46種の依存関係にある2トークン間の関係の種類予測
   - UD: universal dependencies treebank, comprising approx. 2,100 sentences per language (cf. https://universaldependencies.org/treebanks/en_partut/index.html)

- Preposition supersense disambiguation
   3. Preposition's semantic function (PS-fxn, 40種)
   4. Preposition's semantic role (PS-role, 46種)

5. Text classification
   - 文に付与された50種の意味ラベルを予測 (TREC-50)

(1) I was booked **for**/DURATION 2 nights **at**/LOCUS this hotel **in**/TIME Oct 2007 .

(2) I went **to**/GOAL ohm **after**/EXPLANATION⇝TIME reading some **of**/QUANTITY⇝WHOLE the reviews .

(3) It was very upsetting to see this kind **of**/SPECIES behavior especially **in_front_of**/LOCUS **my**/SOCIALREL⇝GESTALT four year_old .

Figure 2: Annotated sentences from the STREUSLE 4.0 corpus, used in the preposition supersense disambiguation task. Prepositions are marked by boldface, immediately followed by their labeled function. If applicable, ⇝ precedes the preposition's labeled role. Figure reproduced from Schneider et al. (2018).

[Liu et al.'21]

実験

Fine-tuningによる精度変化と訓練セットとテストセットのSpatial Similarityの変化

- Fine-tuningの結果、訓練セットとテストセットのSpatial Similarityは常に低下（知見１）
  - Spatial Similarityが計算可能 = linearly separable
  - 同じrepresentationを用い、訓練/テストセットでそれぞれベクトルを作成し、類似度を算出

- 5 tasks × 5 models で唯一、PS-fxn × $\text{BERT}_{\text{small}}$のみfine-tuningにより精度が低下
  - 訓練/テストセット間のSpatial Similarityが大幅に低下
  - それ以外の設定ではいずれも精度は向上
  - <“We conjecture that controlling the divergence between the training and test sets can help ensure that finetuning helps”と書いてあるが訓練セットだけで制御するのは厳しそう>

Task		Acc	Sim
POS	original	94.25	0.96
	tuned	95.43	0.72
DEP	original	92.93	0.93
	tuned	94.48	0.78
PS-fxn	original	86.26	0.82
	tuned	*85.08*	**0.44**
PS-role	original	74.22	0.84
	tuned	74.57	0.54
TREC-50	original	81.32	-
	tuned	89.60	-

Table 2: Fine-tuned performances of $\text{BERT}_{\text{small}}$ based on the last layers. The last column shows the spatial similarity between the training and test set.

Linearity of Representations

- Fine-tuningの結果、クラスタ数は減少
  - 同一ラベルの点をできるだけ少数のクラスタにグルーピング（知見２）
- linearly separableである割合は高い
  - $\text{BERT}_{\text{mini}}$以上のモデルではfine-tuning後はTREC-50以外はlinearly separable
  - $\text{BERT}_{\text{small}}$以上のモデルではfine-tuning前の時点でTREC-50以外はlinearly separable
  - <タスクによる差異が大きそう>

Task		#clusters	is linear	Acc
POS	original	3936	N	90.76
	tuned	20	N	91.67
DEP	original	653	N	86.74
	tuned	46	Y	89.04
PS-fxn	original	402	N	74.14
	tuned	40	Y	74.40
PS-role	original	46	Y	58.38
	tuned	46	Y	60.31
TREC-50	original	399	N	68.12
	tuned	51	N	84.04

Table 3: The linearity of the last layer of $\text{BERT}_{\text{tiny}}$ for each task.

Spatial Structure of Labels

- $\text{BERT}_{\text{base}}$ (=linear)を対象にfine-tuning中のクラスタごとの他のクラスタとの最短距離を調査 ⇒ fine-tuningが進むと最短距離は増加

- Fine-tuningは異なるラベルを表すクラスタを互いに遠ざける（知見３）
  - 右図は距離の変化量上位３ラベルと下位３ラベルのみを表示
  - ただし距離は単調増加ではない ⇒ 潜在的な不安定性を示唆

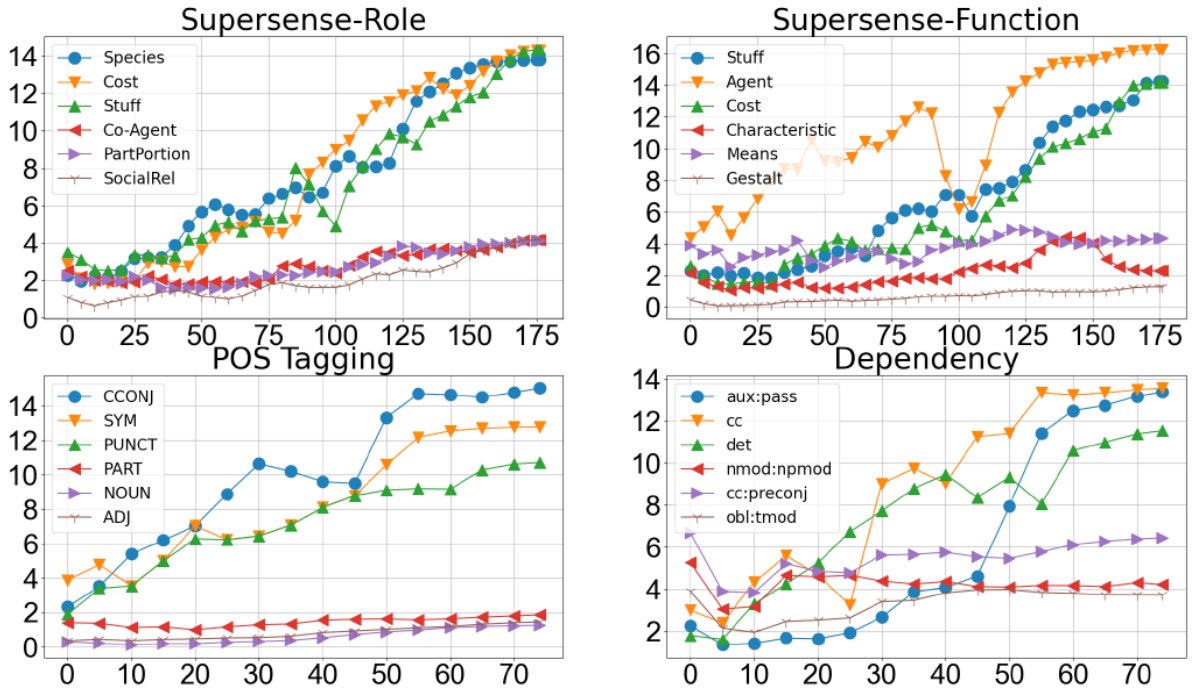


Figure 6: The dynamics of the minimum distance of the three labels where the distance increases the most, and three labels where is increases the least. The horizontal axis is the number of fine-tuning updates; the vertical axis is chosen label's minimum distance to other labels. These results come from the last layer of $\text{BERT}_{\text{base}}$.

クラスタのcentroidの移動

- POS taggingタスクでもっとも近い３種のラベル(ADJ, VERB, NOUN)を対象に調査
- 各クラスタのcentroidのpathをPCAにより２次元にマッピングしたのが右図
- Layer 12を見るとfine-tuning前は非常に近かったcentroidがfine-tuningにより異なる方向に動き離れていくことが確認可能
  - もともと線形分離可能なタスクなのでlossやoptimizerから自明に予測される動きではない

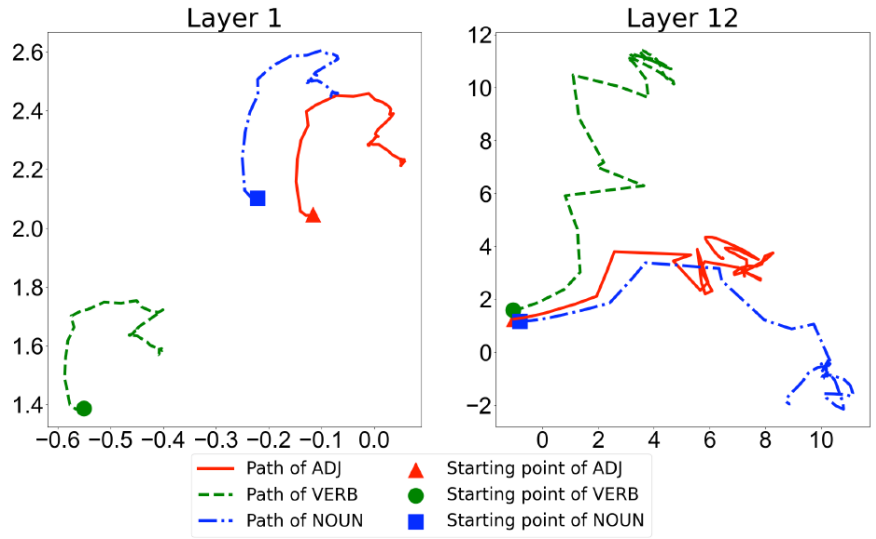


Figure 3: The PCA projection of three closest labels in POS tagging task based on the first (left) and last (right) layer of $\text{BERT}_{\text{base}}$. These lines show the paths of the centroids of each label cluster during the fine-tuning. The markers indicate the starting points. This figure is best seen in color.

Cross-task Fine-tuning

- 異なるタスク間でのfine-tuningの効果を検証
  - Fine-tuningをPS-fxn, PS-role, POSでそれぞれ実施しPS-fnxでテスト
  - PS-roleはPS-fxnは類似タスク
  - POSはPS-fxnの分類対象にすべて同一のラベル（ADP）を付与する敵対的な効果が予測されるタスク
- 類似タスクの場合のみ精度向上
  - クラスタ間の最低距離も類似タスクの場合は増加する傾向（知見４）

fine-tuning	probing	#inc	#dec	average inc	Acc
-	PS-fxn	-	-	-	87.75
PS-fxn	PS-fxn	40	0	5.29	89.58
PS-role	PS-fxn	27	13	1.02	88.53
POS	PS-fxn	0	40	-1.68	83.24

Table 5: Classification performances for PS-fxn task using the last layer of $\text{BERT}_{\text{base}}$ when fine-tuning on different tasks. First row indicates the untuned version. The third and forth column indicate the number of labels whose minimum distance is increased or decreased after fine-tuning. The second last column (average inc) shows the average change of the minimum distance over all the labels. The last column indicates the probing accuracy.

Layer Behavior

- Fine-tuning中の各representationとoriginalのrepresentation間のspatial similarityを計算（右上図）
  - 上位層ほど類似度が低下する傾向
  - ただし、最上位層でも0.5以上と驚くほど大きな値<やや主観的な評価という印象>
  - 上位層も元の構造をある程度保持しており自由に変化している訳ではない（知見５）
- 下位層ではそもそも変化がないのか？
  - fine-tuning前後の全クラスタのcentroidの変化をPCAにより２次元に射影（右下図）
  - 変化はあるが同方向に小さく変化する傾向

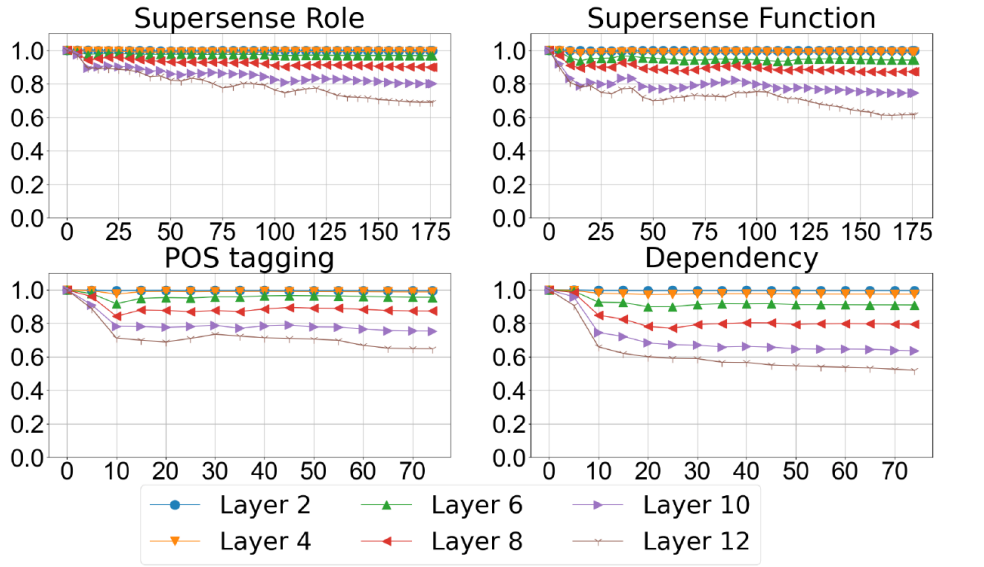


Figure 4: Dynamics of spatial similarity during the fine-tuning process based on $\text{BERT}_{\text{base}}$.

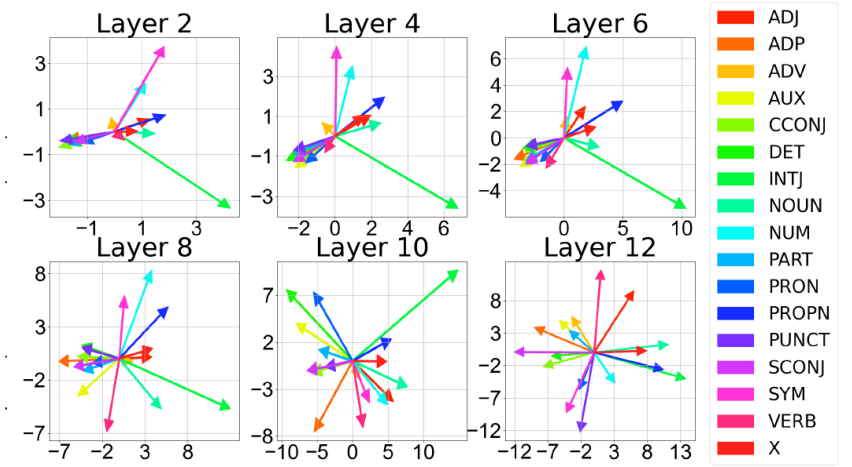


Figure 5: The PCA projection of the difference vector between the centroids of labels before and after fine-tuning based on POS tagging task and $\text{BERT}_{\text{base}}$.

まとめと所感

- まとめ: Fine-tuningによるBERTの変化を調査 / 得られた主な知見は以下
  1. Fine-tuning introduces a divergence between training and test sets, which is not severe enough to hurt generalization in most cases
  2. Fine-tuning adjusts a representation by grouping points with the same label into a small number of clusters (ideally one)
  3. Fine-tuning pushes the clusters of points representing different labels away from each other, thus introducing large separating regions between labels
  4. Fine-tuning for related tasks can also provide useful signal for the target task by altering the distances between clusters representing different labels
  5. Fine-tuning does not change the higher layers arbitrarily <やや主観的？>
- 所感
  - <実際の空間上の変化(cf. centroid)ではなく、DIRECTPROBEを用いて構築される、各ラベルに対応するクラスタの相対的な位置関係に基づき分析しているのが面白い>
  - <Fine-tuning前の時点でlinearなタスクが主なので汎用的な結論が導けるのかはやや疑問>
  - <線形分離が難しいタスク、より大規模なモデルに関する考察、が今後の課題？>

付録: Table 7: A complete table of the probing results of five representations on five tasks

Representations	Task		Acc	Std	Best Layer Size	#Cluster	is Linear	Similarity
$BERT_{tiny}$	POS	original	90.76	0.24	(256, 64)	3936	N	-
		fine-tuned	91.67	0.29	(64, 64)	20	N	-
	DEP	original	86.74	0.22	(256, 256)	653	N	-
		fine-tuned	89.04	0.20	(256, 256)	46	Y	0.88
	PS-fxn	original	74.14	1.42	(256, 256)	402	N	-
		fine-tuned	74.40	0.68	(256, 128)	40	Y	0.72
	PS-role	original	58.38	0.78	(256, 64)	46	Y	0.76
		fine-tuned	60.31	0.29	(64, 64)	46	Y	0.70
	TREC-50	original	68.12	0.82	(256, 256)	399	N	-
		fine-tuned	84.04	0.93	(256, 256)	51	N	-
$BERT_{mini}$	POS	original	93.81	0.10	(256, 32)	2429	N	-
		fine-tuned	94.91	0.03	(256, 32)	17	Y	0.70
	DEP	original	91.82	0.09	(256, 128)	46	Y	0.93
		fine-tuned	93.55	0.07	(256, 128)	46	Y	0.86
	PS-fxn	original	82.45	1.07	(256, 256)	40	Y	0.77
		fine-tuned	84.25	0.39	(256, 128)	40	Y	0.53
	PS-role	original	68.05	1.08	(256, 256)	46	Y	0.81
		fine-tuned	71.90	1.06	(256, 64)	46	Y	0.59
	TREC-50	original	74.12	1.25	(256, 256)	127	N	-
		fine-tuned	88.36	0.50	(64, 32)	52	N	-

$BERT_{small}$	POS	original	94.26	0.13	(256, 32)	17	Y	0.96
		fine-tuned	95.43	0.06	(128, 64)	17	Y	0.72
	DEP	original	92.93	0.14	(256, 64)	46	Y	0.93
		fine-tuned	94.48	0.14	(256, 64)	46	Y	0.78
	PS-fxn	original	86.26	0.54	(256, 256)	40	Y	0.82
		fine-tuned	85.08	0.35	(256, 256)	40	Y	0.44
	PS-role	original	74.22	1.03	(256, 256)	46	Y	0.84
		fine-tuned	74.57	0.61	(128, 128)	46	Y	0.54
	TREC-50	original	81.32	0.61	(256, 128)	113	N	-
		fine-tuned	89.60	0.22	(256, 64)	51	N	-
$BERT_{medium}$	POS	original	94.40	0.08	(256, 128)	17	Y	0.97
		fine-tuned	95.56	0.05	(64, 32)	17	Y	0.67
	DEP	original	92.54	0.14	(256, 256)	46	Y	0.94
		fine-tuned	94.76	0.20	(128, 128)	46	Y	0.79
	PS-fxn	original	86.56	0.41	(256, 128)	40	Y	0.80
		fine-tuned	88.45	0.45	(128, 256)	40	Y	0.59
	PS-role	original	76.28	1.00	(256, 32)	46	Y	0.83
		fine-tuned	78.86	0.58	(128, 128)	46	Y	0.58
	TREC-50	original	80.68	1.16	(256, 64)	110	N	-
		fine-tuned	89.80	0.33	(32, 64)	52	N	-
$BERT_{base}$	POS	original	93.39	0.31	(256, 128)	17	Y	0.97
		fine-tuned	95.68	0.02	(128, 64)	17	Y	0.70
	DEP	original	89.39	0.08	(256, 128)	46	Y	0.92
		fine-tuned	94.76	0.05	(64, 256)	46	Y	0.76
	PS-fxn	original	87.75	0.41	(256, 128)	40	Y	0.84
		fine-tuned	89.58	0.67	(32, 256)	40	Y	0.57
	PS-role	original	74.49	0.84	(256, 128)	46	Y	0.82
		fine-tuned	81.14	0.26	(256, 128)	46	Y	0.52
	TREC-50	original	85.24	0.85	(256, 128)	162	N	-
		fine-tuned	90.36	0.32	(64, 32)	51	N	-