CPE393 : CODING AI PRESENTATION

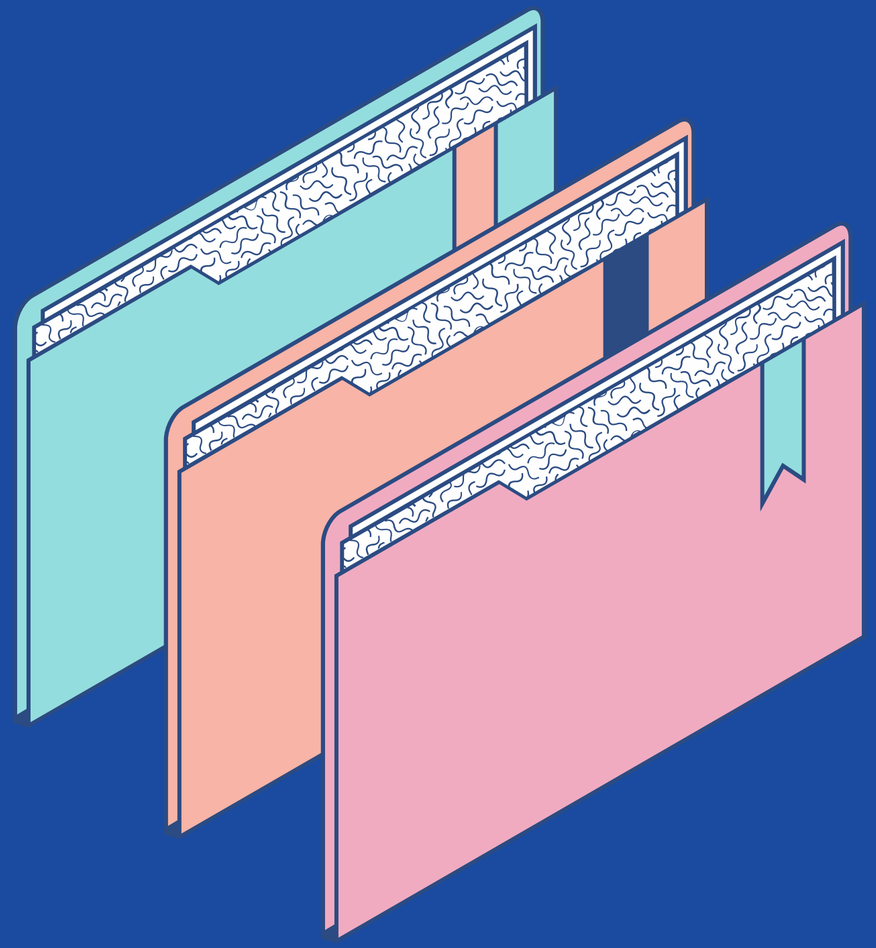
BERT

Bidirectional Encoder Representations from Transformers

Group members

นาย กิติพัฒน์ เรืองอมรวัฒน์
63070501006

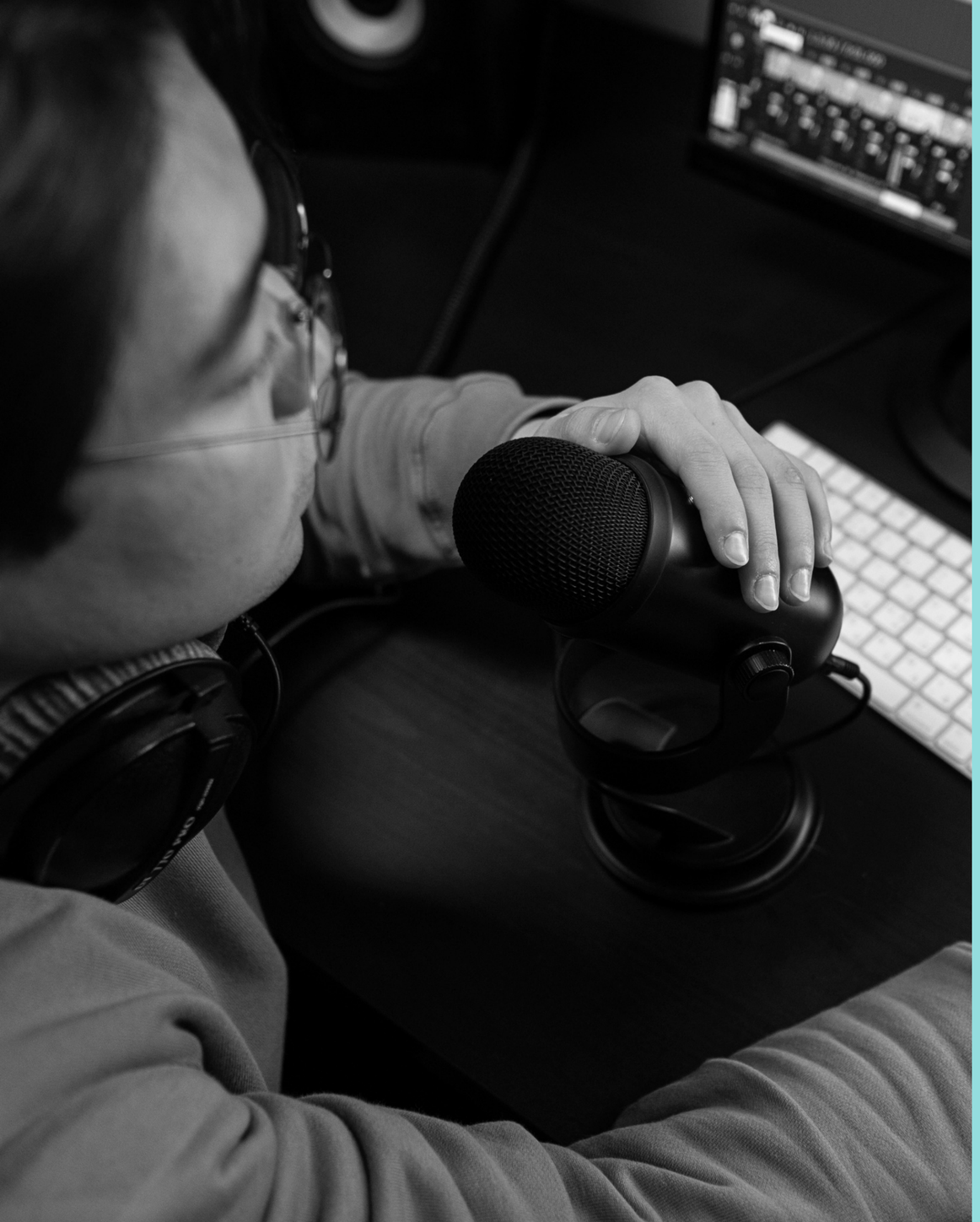
นาย สัณหณัฐ พรมจรรย์
63070501069



Key topics

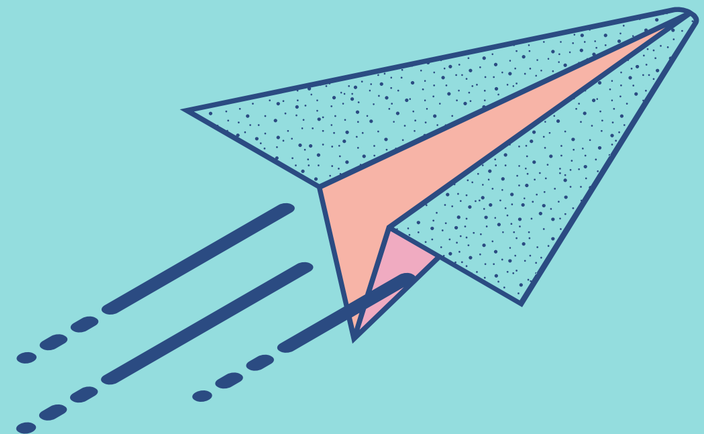
DISCUSSED IN THIS PRESENTATION

- What is BERT?
- The Intro of BERT Research
- Theory
- Experiment & Results
- Ablation Studies
- Conclusion



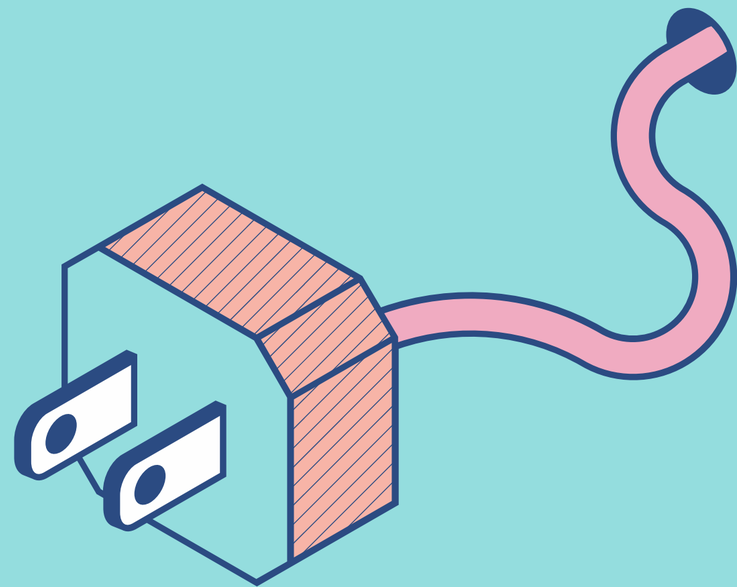
What is BERT?

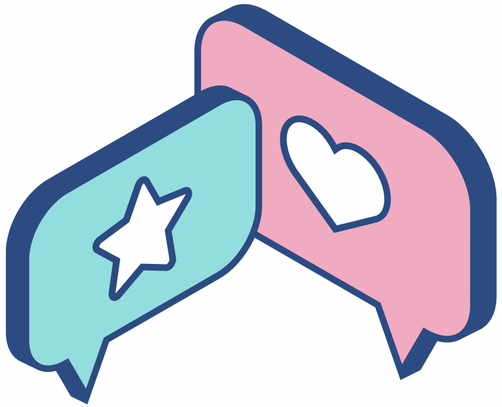
BERT ซึ่งย่อมาจาก Bidirectional Encoder Representations from Transformers ซึ่ง BERT เป็นแบบจำลอง การแสดงภาษาที่ฝึกการแสดงแบบ deep bidirectional



The Intro of BERT Research

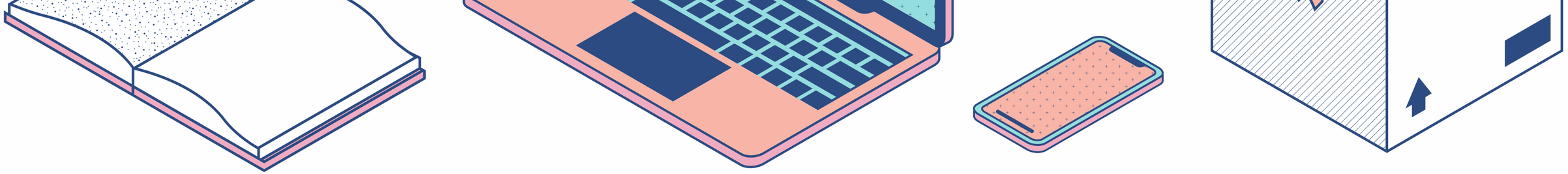
- Natural Language Processing tasks
- 2 strategies: feature-based, fine-tuning
- Unidirectional & Bidirectional
- Masked Language Model (MLM)





Natural Language Processing tasks

- Language model pre-training ได้แสดงให้เห็นถึงประสิทธิภาพในการพัฒนางาน natural language processing
- งาน NLP เช่น การตีความหมายหรือการถอดความของภาษาทั่วไป
- มีจุดมุ่งหมายเพื่อ predict ความสัมพันธ์ระหว่างประโยคโดยการวิเคราะห์แบบองค์รวม ตลอดจนงาน token-level เช่น named entity recognition (การกำหนดเอกลักษณ์หรือนิพจน์ของคำ), Q&A
- จำเป็นต้องมีความละเอียดระดับ token-level

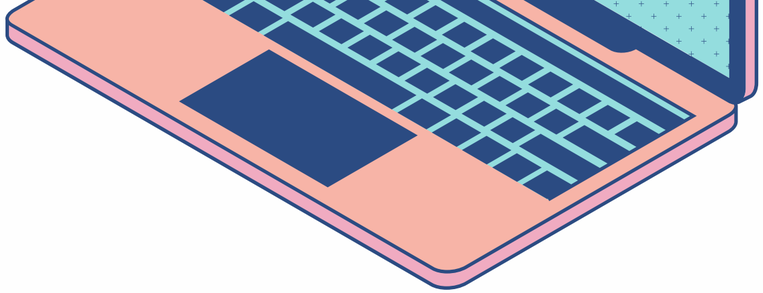


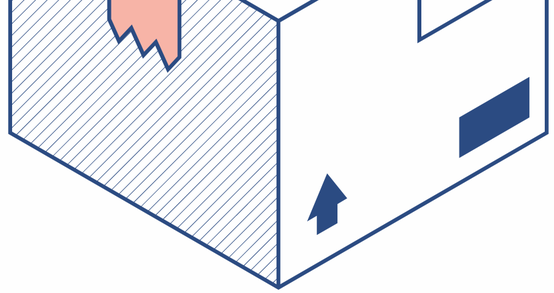
Feature-based

- การใช้ feature ที่มีอยู่โดยนำ pre-trained โมเดลมาใช้ในการทำงานต่างๆ
- ใช้ architecture ที่เฉพาะเจาะจงตามงานที่ต้องทำ และรวมส่วนประมวลผลที่ถูก pre-trained ไว้มาเป็นฟีเจอร์เพิ่มเติมเข้าไปในโมเดลนิวรอนที่ใช้สำหรับงานนั้น ๆ เช่น ELMo
- เมื่อใช้วิธีนี้ ข้อมูลฟีเจอร์ที่ถูก pre-trained ไว้จะเสริมประสิทธิภาพให้กับโมเดลในการเข้าใจและทำงานกับงานต่างๆ เช่น การจำแนกข้อความ การระบุบุคคลและสถานที่ หรือการตอบคำถาม

Fine-tuning

- คือ กระบวนการที่ใช้ปรับแต่งโมเดลที่ถูก pre-trained ไว้เพื่อให้มันมีประสิทธิภาพในงานที่เฉพาะเจาะจงมากขึ้น
- เช่น Generative Pre-trained Transformer (OpenAI GPT)
- สามารถ train เกี่ยวกับงาน downstream ต่างๆ โดยเพียงแค่ปรับแต่งพารามิเตอร์ที่ได้รับการ pre-trained ทั้งหมดอย่างละเอียด



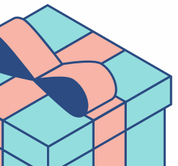


Unidirectional

- ข้อมูลไหลไปในทิศทางเดียวจากซ้ายไปขวาหรือขวาไปซ้าย
- เช่น ในโมเดล unidirectional language จากซ้ายไปขวา โมเดลจะประมวลผลคำจากซ้ายไปขวาในประโยค โดยจะใช้เฉพาะคำที่มาก่อนคำปัจจุบันเพื่อคาดเดาคำถัดไป
- คำนวณง่าย ไม่ซับซ้อน เนื่องจากโมเดลจะประมวลผลข้อมูลตามลำดับ ทีละโทเค็นจึงทำให้ train ได้ง่าย

Bidirectional

- ในทางกลับกัน ข้อมูลจะถูกพิจารณาจากทั้งสองทิศทาง ทั้งข้อมูลที่อยู่ก่อนและข้อมูลที่อยู่หลังของคำในประโยค จะถูกนำมาใช้ในการทำนายคำปัจจุบัน
- เช่น การแสดงคำหนึ่งคำขึ้นอยู่กับคำที่อยู่ข้างหน้าและคำที่อยู่ข้างหลังในประโยคต่างกับแบบ unidirectional ที่จะพิจารณาเพียงทางใดทางหนึ่งในการแสดงคำต่อไปของประโยค
- สามารภจับต้นทางของบริบทได้มากขึ้นเนื่องจากมันพิจารณาทั้งข้อมูลที่เป็นคำก่อนหน้าและข้อมูลที่เป็นคำต่อไป สิ่งนี้มีความสำคัญสำหรับงานที่ต้องการทำความเข้าใจความหมายของประโยค ที่บ่งบอกถึงความหมายของทั้งบริบท

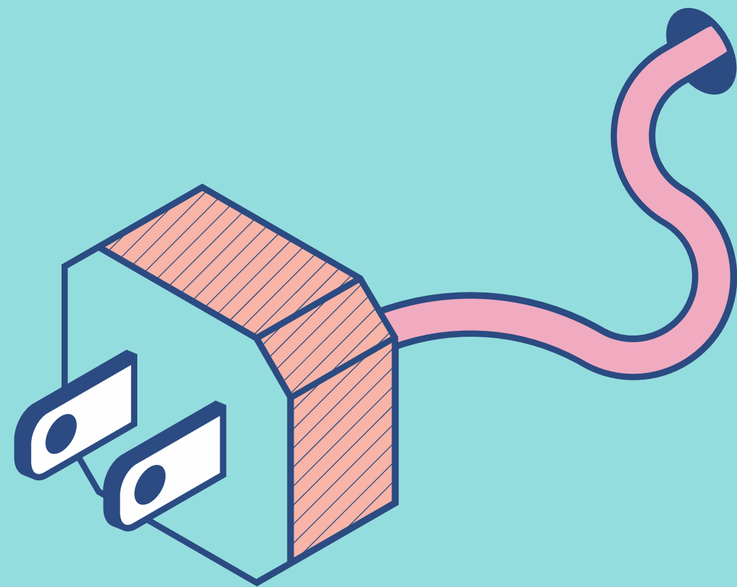


Masked Language Model

- หมายถึง "การทำนายคำที่หายไป"
- MLM นั้นจะทำการสุ่มในการ masks บาง token จาก input โดยมีวัตถุประสงค์เพื่อ predict id ดั้งเดิมของคำศัพท์ที่ masked ไว้โดยอ้างอิงตามบริบทเท่านั้น
- ไม่เหมือนกับโมเดลที่เป็น left-to-right language MLM นั้นจะเชื่อมบริบททั้งทิศทางซ้ายและขวาซึ่งก็ คือ left-to-right และ right-to-left
- ช่วยทำให้สามารถ pre-train deep bidirectional transformer ได้แล้วยังใช้ในการทำนายประโยคถัดไป

BERT Theory

- Pre-training general language
- BERT



Pre-training general language

Unsupervised feature-based

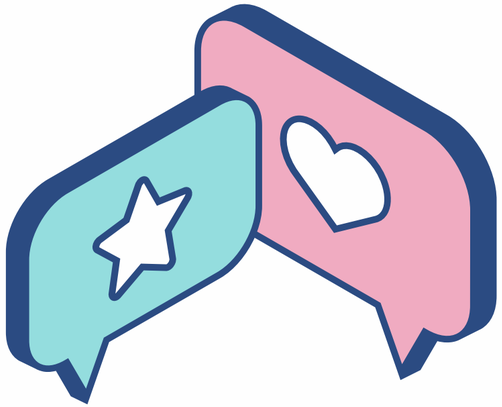
เป็นการเรียนรู้การแสดงคำทั้งแบบวิธี non-neural และ neural ด้วย pre-trained word embeddings แสดงให้เห็นถึงการพัฒนา scratch learned embeddings ผ่านการสร้าง language model และการแยกแยะความถูกต้องจากคำที่ไม่ถูกต้องซึ่งยังถูกใช้กับงานที่มีรายละเอียดที่ซับซ้อน

Unsupervised Fine-tuning

มีการ pre-trained word embeddings พารามิเตอร์จาก unlabeled text เช่นเดียวกันกับ feature-based จากนั้นได้นำไปใช้กับ sentence หรือ document encoders ซึ่งการสร้างโทเคนตามบริบทในการแสดงนั้นก็ถูก pre-trained จาก unlabeled text และได้รับการ fine-tuning ใน supervised downstream task

Transfer Learning from Supervised Data

ใช้ชุดข้อมูลขนาดใหญ่ที่มีประสิทธิภาพในการปรับปรุงประสิทธิภาพเป็น natural language inference(การอนุมานภาษาธรรมชาติ) และ machine translation(การแปลด้วยเครื่อง)



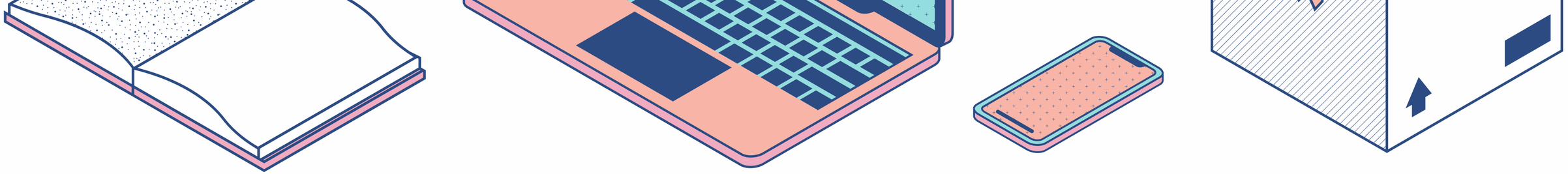
BERT

- เป็น framework ที่ประกอบด้วย 2 ขั้นตอน คือการ **pre-training** และการ **fine-tuning**

- **คุณลักษณะที่โดดเด่นของ BERT** คือมี unified architecture ใน task ต่าง ๆ โดยมีความแตกต่างน้อยที่สุด ระหว่าง architecture ที่ pre-trained และ final downstream

- **Model Architecture** คือ multi-layer bidirectional Transformer encoder based

- **Input/Output Representations** เพื่อให้ BERT จัดการ downstream tasks ที่หลากหลาย การแสดงอินพุตจึงสามารถแสดงทั้งประโยคเดี่ยวหรือคู่ได้อย่างชัดเจน (เช่น คำถาม คำตอบ) ในลำดับโทเคนเดียว



Pre-training BERT (Two unsupervised task)

Masked LM

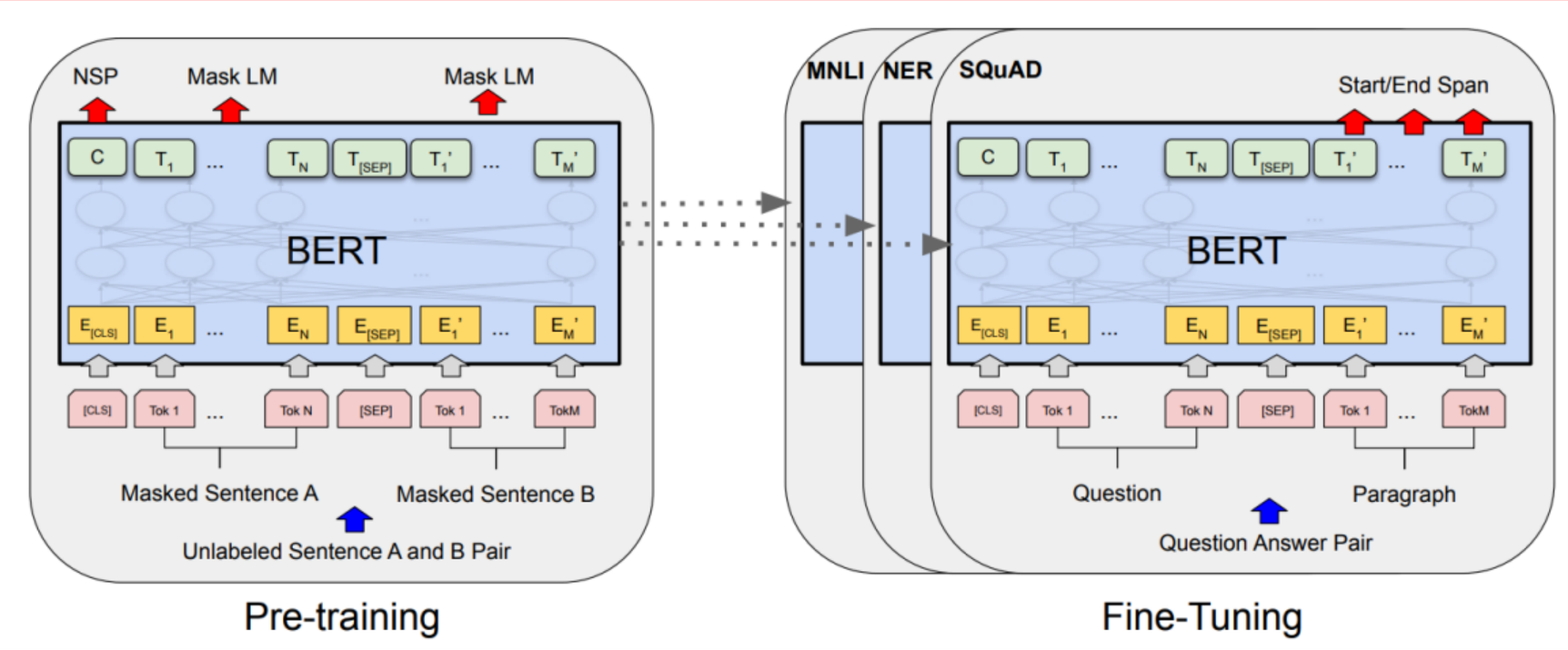
- เกี่ยวข้องกับ mask แบบสุ่ม 15% ของโทเคนอินพุตและทำนายโทเค็นที่ mask เหล่านั้นโดยใช้เอาต์พุต SoftMax
- การใช้ deep bidirectional model ใน BERT ช่วยให้สามารถแสดงข้อมูลได้มีประสิทธิภาพมากขึ้นเมื่อเปรียบเทียบกับโมเดลแบบดั้งเดิม
- ข้อเสียคือ โทเค็น [MASK] ถูกใช้ระหว่าง pre-training แต่ไม่ปรากฏระหว่าง fine-tuning ทำให้เกิดความไม่ตรงกัน

Next Sentence Prediction (NSP)

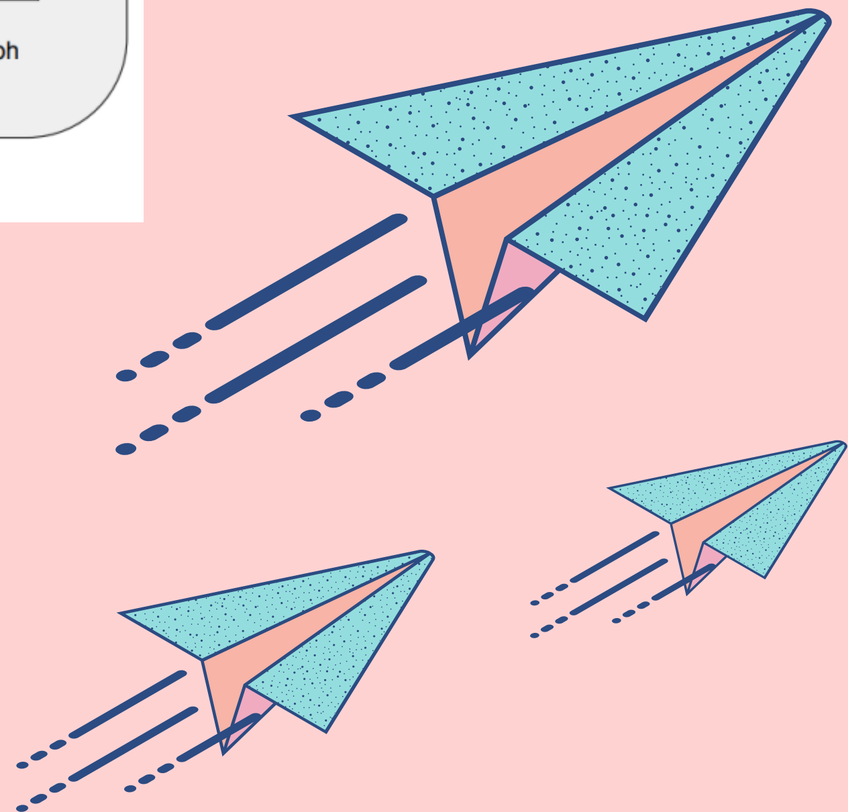
- เป็น binarized task ที่สามารถสร้างได้เพียงเล็กน้อย จากคลังข้อมูลภาษาเดียว
- โดย 50% ของประโยคถัดไปจะตามหลังประโยคปัจจุบัน และ 50% ของประโยคจะถูกเลือกแบบสุ่มจากคลังข้อมูล

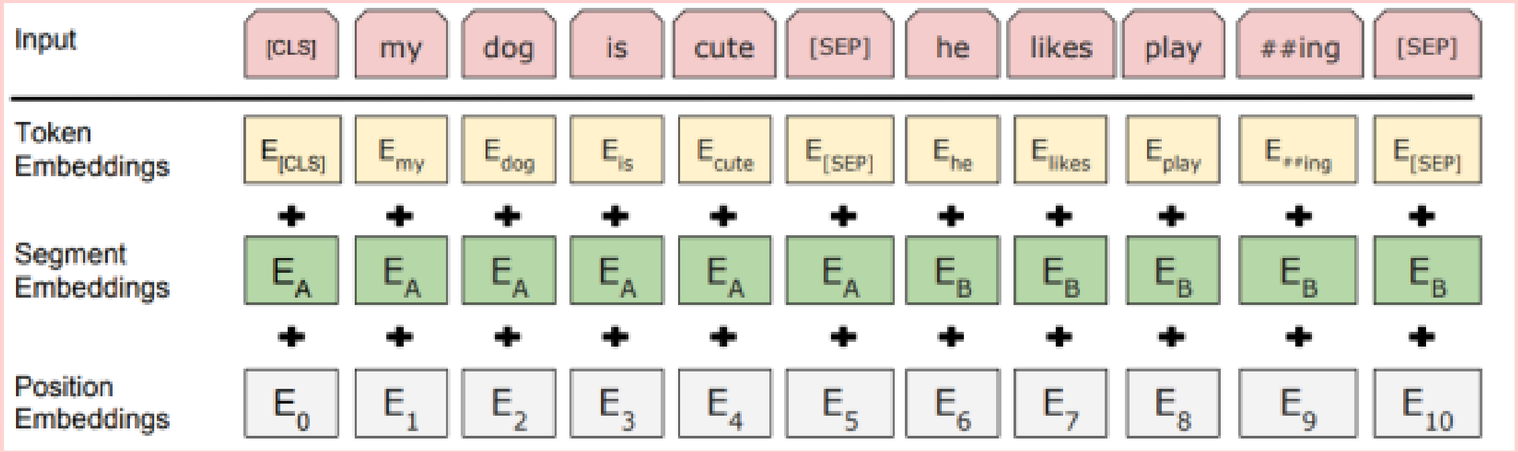
Fine-tuning BERT

- Fine-tuning ทำปกติไปเลย เนื่องจากใช้ self-attention mechanism ใน Transformer
- ช่วยให้ BERT สามารถสร้างโมเดล downstream tasks จำนวนมากได้ ไม่ว่าจะเกี่ยวข้องกับข้อความเดี่ยวหรือคู่ข้อความ
- เนื่องจาก encoding คู่ข้อความที่ต่อกันเข้ากับ self-attention นั้นรวม bidirectional cross-attention ระหว่างสองประโยคอย่างมีประสิทธิภาพ

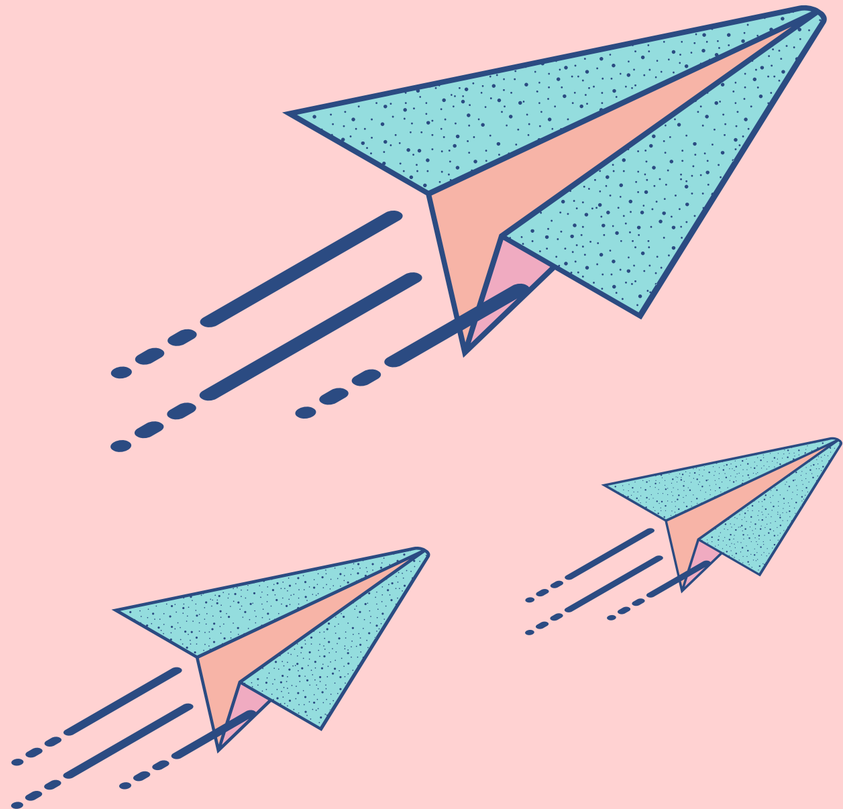


ขั้นตอน pre-training และ fine-tuning โดยรวมสำหรับ BERT นอกเหนือจากเลเยอร์เอาต์พุตแล้ว architecture เดียวกันนี้ยังใช้ทั้งใน pre-training และ fine-tuning พารามิเตอร์โมเดลที่ได้รับ pre-trained เดียวกันนี้ใช้เพื่อเริ่มต้นโมเดลสำหรับ downstream tasks ที่แตกต่างกัน และในระหว่าง fine-tuning พารามิเตอร์ทั้งหมดจะได้รับ fine-tuning



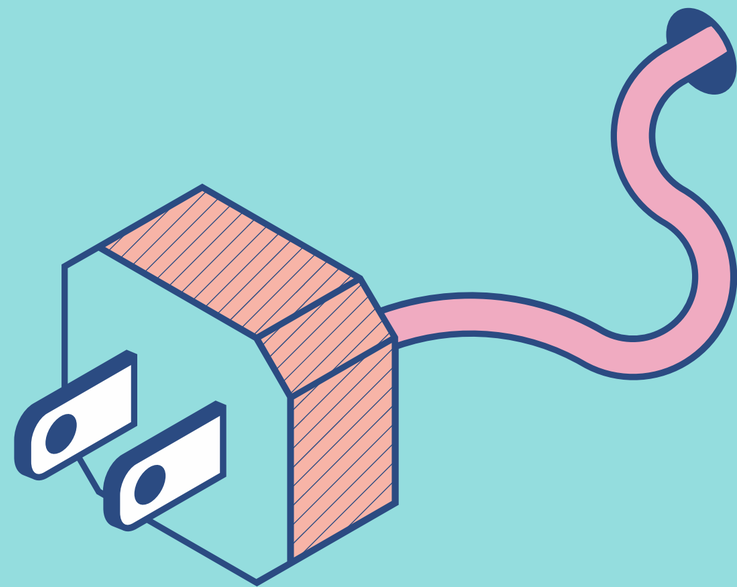


การแสดงอินพุตจะถูกสร้างขึ้นโดยการรวม Token, segment, และ position embeddings ที่สอดคล้องกัน



Experiment & Results

- GLUE
- SQuAD v1.1
- SQuAD v2.0
- SWAG



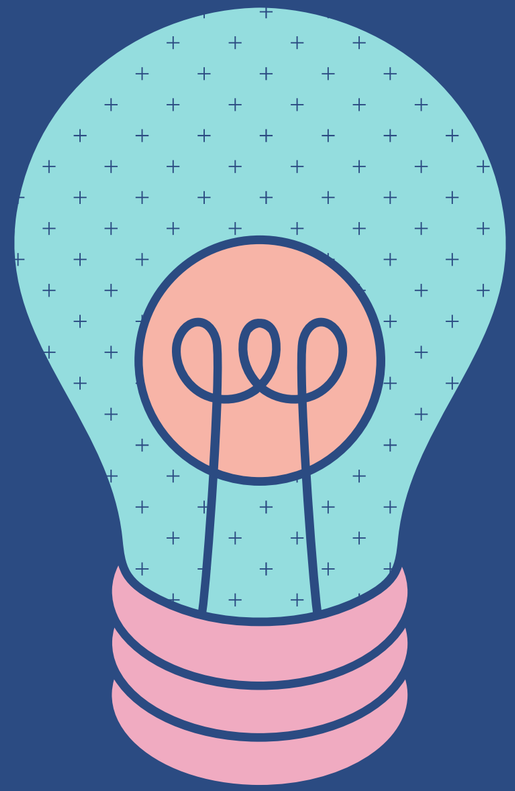
GLUE

- The General Language Understanding Evaluation (GLUE) คือเกณฑ์มาตรฐานที่มีคอลเล็คชันที่หลากหลายของงาน การทำความเข้าใจ natural language

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

- นักวิจัยใช้ batch size 32 และ fine-tune 3 epoch สำหรับงาน GLUE ทั้งหมด ในแต่ละงานนักวิจัยได้ใช้ learning rate ที่ดีที่สุด (ระหว่าง 5e-5, 4e-5, 3e-5 และ 2e-5) ในการทดสอบ GLUE

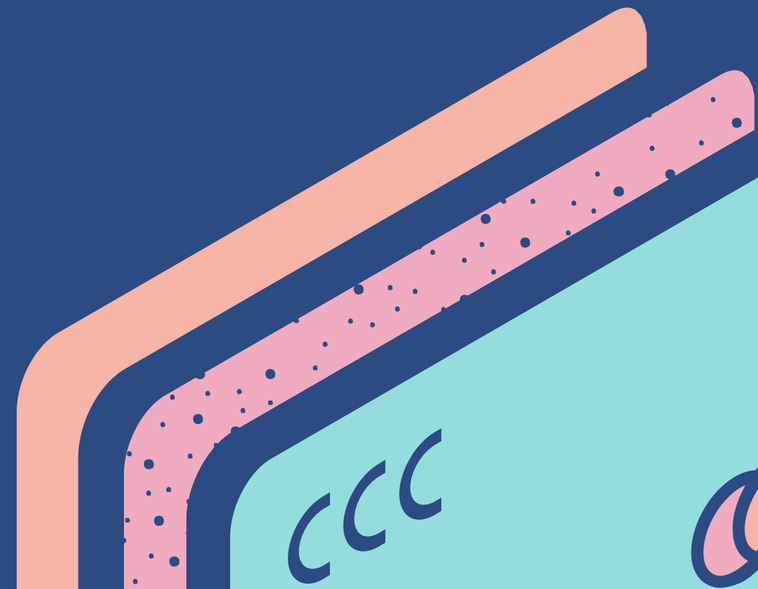
- จากผลลัพธ์ที่ได้ทั้ง BERT(base) และ BERT(large) มีประสิทธิภาพเหนือกว่าทุก systems ในทุกๆด้าน เมื่อเทียบกับเทคโนโลยีรุ่นก่อน และได้พบว่าทั้ง BERT(base) และ BERT(large) ยังทำงานได้ดีต่อให้ training data จะน้อยก็ตาม



SQuAD v1.1

- The Stanford Question Answering Dataset (SQuAD v1.1) คือ ชุดของคู่คำถาม/คำตอบที่มาจากฝูงชน 100,000 คู่
- เมื่อมีคำถามและข้อความจากวิกิพีเดียที่มีคำตอบ โดยการทดสอบคือจะต้องคาดเดาคำตอบในช่องว่าง
- โดยนักวิจัยได้ fine-tune ใน 3 epoch ด้วย learning rate 5e-5 และ batch size 32
- ระบบที่ทำงานดีที่สุดของเรามีประสิทธิภาพดีกว่าระบบที่อยู่บน leaderboard อันดับหนึ่งขึ้นมาถึง +1.5 F1 เมื่อใช้กระบวนการอะสมข้อมูล (ensembling) และ +1.3 F1 เมื่อใช้เป็นระบบเดียวเท่านั้น ในความเป็นจริง, โมเดล BERT มีประสิทธิภาพดีกว่าระบบ ensemble ที่ดีที่สุดในเชิงคะแนน F1 โดยไม่ใช้ข้อมูลจาก TriviaQA ในกระบวนการปรับแต่ง, เสียเพียง 0.1-0.4 F1 เท่านั้น แต่ยังมีประสิทธิภาพที่ดีกว่าระบบที่มีอยู่ทั้งหมดอย่างชัดเจน

System	Dev		Test	
	EM	F1	EM	F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
Published				
BiDAF+ELMo (Single)	-	85.6	-	85.8
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**



SQuAD v2.0

- ในการทดสอบ SQuAD v2.0 จะเพิ่มการกำหนดปัญหาจาก SQuAD 1.1 โดยอนุญาตให้ไม่มีคำตอบสั้นๆ ในย่อหน้าที่ให้มา ทำให้สมจริงยิ่งขึ้น
- ผลลัพธ์นั้นแสดงในตาราง เปรียบเทียบกับโมเดลก่อนๆ ไม่รวมระบบที่ใช้ BERT เป็นหนึ่งในองค์ประกอบจะเห็นได้ว่าค่า F1 เพิ่มขึ้น +5.1 จากระบบที่ดีที่สุดก่อนหน้านี้
- ใน fine-tuned ที่ 2 epoch นักวิจัยใช้ learning rate 5e-5 และ batch size 48

System	Dev		Test	
	EM	F1	EM	F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	86.3	89.0	86.9	89.5
#1 Single - MIR-MRC (F-Net)	-	-	74.8	78.0
#2 Single - nlnet	-	-	74.2	77.1
Published				
unet (Ensemble)	-	-	71.4	74.9
SLQA+ (Single)	-		71.4	74.4
Ours				
$\text{BERT}_{\text{LARGE}}$ (Single)	78.7	81.9	80.0	83.1

SWAG

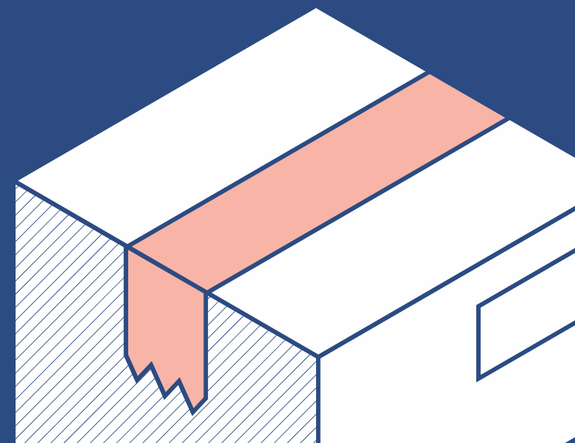
- The Situations With Adversarial Generations(SWAG) เป็นชุดข้อมูลที่มีตัวอย่างการเติมประโยคคู่สมบูรณ์ 113,000 ตัวอย่างที่ประเมินการอนุมานสามัญสำนึกที่มีพื้นฐาน

- เมื่อพิจารณาจากประโยคแล้ว การทดสอบคือเลือกประโยคต่อเนื่องที่เป็นไปได้มากที่สุดจากสี่ตัวเลือก

- ในระหว่างการปรับแต่งชุดข้อมูล SWAG อย่างละเอียด ลำดับอินพุตสี่ลำดับจะถูกสร้างขึ้นโดยการรวมประโยคที่กำหนดเข้ากับความต่อเนื่องที่เป็นไปได้ และมีการใช้พารามิเตอร์เฉพาะงานเพื่อคำนวณคะแนนสำหรับแต่ละตัวเลือก

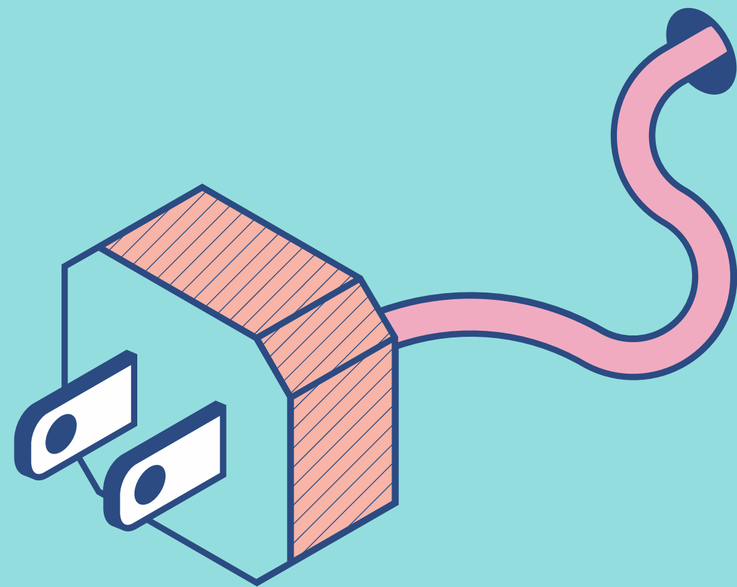
- นักวิจัย fine-tune โมเดลใน 3 epoch ด้วย learning rate 2e-5 และ batch size 16 ผลลัพธ์ถูกแสดงที่ตาราง BERT(large) มีประสิทธิภาพเหนือกว่าระบบ the authors' baseline ESIM+ELMo +27.1% และ OpenAI GPT 8.3 %

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
OpenAI GPT	-	78.0
$\text{BERT}_{\text{BASE}}$	81.6	-
$\text{BERT}_{\text{LARGE}}$	**86.6**	**86.3**
Human (expert)†	-	85.0
Human (5 annotations)†	-	88.0



Ablation studies

- Effect of Pre-training Tasks
- Effect of Model size
- Feature-based Approach with BERT



Conclusion

- การ transfer learning ด้วย LM ได้แสดงให้เห็นถึงการ pretraining ที่สมบูรณ์และ unsupervised เป็นส่วนสำคัญของระบบการทำความเข้าใจภาษาต่างๆ ผลลัพธ์เหล่านี้ช่วยให้งานที่ใช้ทรัพยากรต่ำได้รับประโยชน์จาก unidirectional architecture เชิงลึก
- ผลงานที่สำคัญของงานวิจัยนี้คือการสรุปการค้นพบเหล่านี้ให้เป็นแบบ bidirectional architecture เชิงลึกซึ่งช่วยให้โมเดลที่ได้รับการ pre-trained แบบเดียวกันสามารถจัดการกับงาน NLP ในวงกว้างได้สำเร็จ

THANK YOU