ΤΕΛΙΚΗ ΑΝΑΦΟΡΑ

της εξαμηνιαίας εργασίας με τίτλο:

«REINFORCEMENT LEARNING AND DISTILLATION»

για το μάθημα

Αναγνώριση Προτύπων

Μέλη της ομάδας:

1) Δωροθέα Κουμίδου 03119712

2) Γιώργος Γκοτζιάς 03119047

3) Γιώργος Χαραλάμπους 03119706

ΕΙΣΑΓΩΓΗ

Η παρούσα μελέτη πραγματεύεται με την μεταφορά γνώσης από προ-εκπαιδευμένα μοντέλα όπως το BERT σε μικρότερου όγκου μοντέλα Distilbert. Στην συνέχεια, προβάλλεται η λειτουργία ενός άλλου PLM, ELECTRA, και ερευνάται η δυνατότητα εφαρμογής της διαδικασία «distillation» στο συγκεκριμένο μοντέλο.

ΔΙΑΤΥΠΩΣΗ ΕΡΕΥΝΗΤΙΚΟΥ ΕΡΩΤΗΜΑΤΟΣ

Το ερευνητικό ερώτημα που αποτέλεσε έναυσμα της παρούσας μελέτης είναι αν είναι δυνατή η εύρεση μιας διαδικασίας distillation και η εφαρμογή της στο μοντέλο ELECTRA. Σε περίπτωση που δεν είναι εφικτό, διερευνείται το knowledge distillation του BERT σε ένα μικρότερο υποσύνολο δεδομένων.

ΠΕΡΙΓΡΑΦΗ ΜΕΘΟΔΟΛΟΓΙΑΣ

Σε μια πρώτη επαφή με το θέμα, απαραίτητη προϋπόθεση είναι η βαθιά κατανόηση της λειτουργίας του distilBert. Ο όρος knowledge distillation(KD) αναφέρεται στη μεταφορά γνώσης από ένα προ-εκπαιδευμένο μοντέλο μεγάλου όγκου δεδομένων, όπως το BERT, σε ένα μικρότερο μοντέλο, το οποίο βασίζεται στην αρχιτεκτονική του μεγάλου.

Tokenization

Στη συνέχεια, έγινε μια βιβλιογραφική έρευνα για τη κατανόηση της λειτουργίας του BERT (Εικόνα 1), το οποίο είναι masked language model. Τέτοια μοντέλα δέχονται σαν είσοδο μια πρόταση την οποία με χρήση tokens (CLS, mask, SEP) την κωδικοποιούν. Η λειτουργία των tokens αφορά:

CLS: το σημείο αναγνώρισης της πρότασης και τοποθετείται στην αρχή της

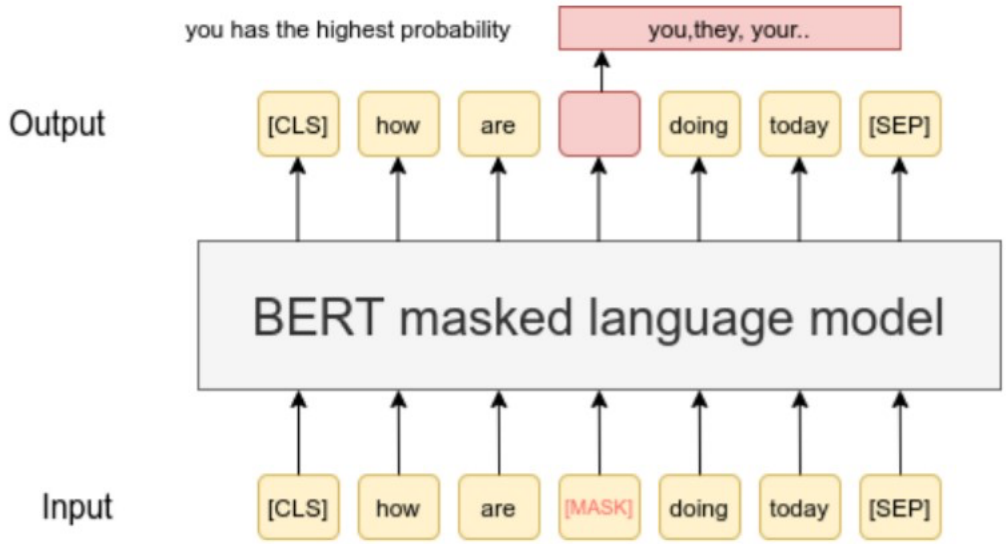
Mask: την τυχαία επιλογή λέξεων της πρότασης, περίπου 15%, οι οποίες θα πρέπει να αντικατασταθούν κατά την κωδικοποίηση ως token mask και προορίζονται να προβλεφθούν

SEP: να διαχωρίσει τις προτάσεις εισόδου και τοποθετείται στο τέλος της πρότασης

Bert

Εφόσον έγινε αντιληπτή η ορολογία των tokens, παρουσιάζεται η λειτουργία του μοντέλου BERT. Η είσοδος είναι μια πρόταση ή περισσότερες, που έχουν ήδη μετατραπεί σε tokens. Οι λέξεις που επιλέχθηκαν ως mask στο BERT αντικαθίστανται με βάση την προ-εκπαιδευμένη γνώση του μοντέλου. Η αντικατάσταση επιλέγεται με κριτήριο μιας κατανομή

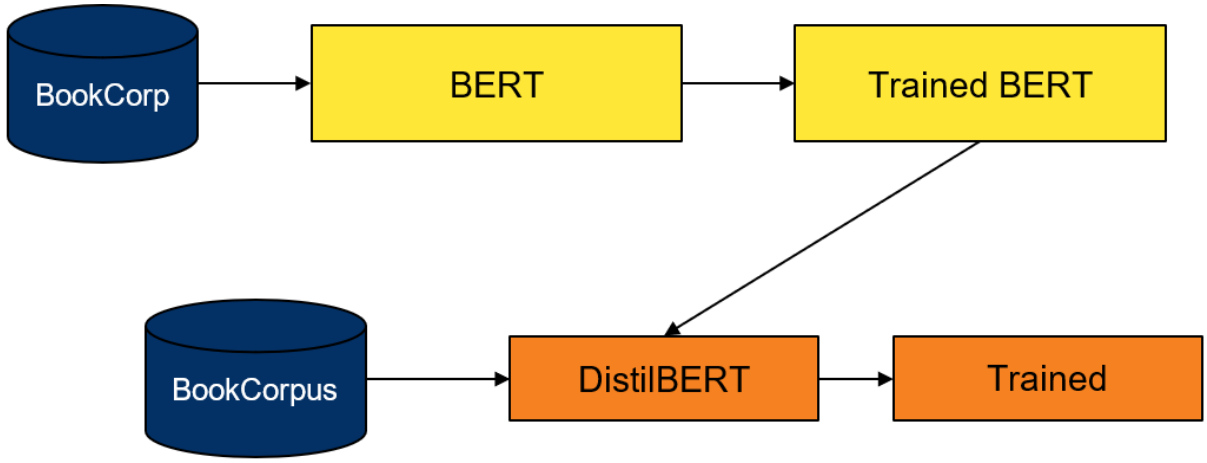
λέξεων με τις αντίστοιχες πιθανότητες εμφάνισης τους, βάρη, την οποία παίρνει έτοιμη από την προ-εκπαίδευση. Τελικά το mask token εμφανίζεται στην έξοδο ως η λέξη που είχε την μεγαλύτερη πιθανότητα. Τα υπόλοιπα tokens μένουν αναλλοίωτα αφού περιορίζονται απλά στην σηματοδότηση της αρχής και του τέλους της πρότασης. Η παραπάνω διαδικασία χαρακτηρίζεται ως generator επειδή η έξοδος επιλέγεται από μια κατανομή πιθανοτήτων και είναι συνεχής.



Εικόνα 1: Αρχιτεκτονική BERT

DistilBert

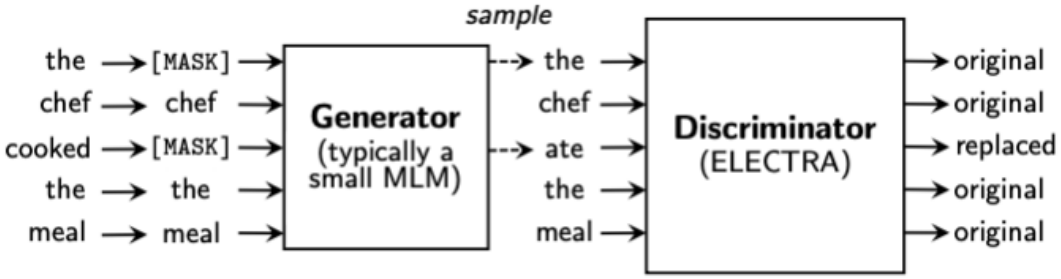
Η λειτουργία του distilBERT ακολουθεί την γενική αρχιτεκτονική του μοντέλου BERT. Ορίζεται μια σχέση δασκάλου- μαθητή, όπου ο δάσκαλος είναι ένα BERT με μεγάλο όγκο δεδομένων και ο μαθητής είναι το distilBERT με πολύ μικρότερο όγκο δεδομένων. Ουσιαστικά γίνεται μεταφορά γνώσης από τον δάσκαλο στο μαθητή , ώστε ο μαθητής να είναι σε θέση να επιλέξει την σωστή αντικατάσταση του mask, σε λιγότερο χρόνο εκπαίδευσης με ακρίβεια ίση ή μεγαλύτερη του δασκάλου. Η εικόνα 2 απεικονίζει τη σχέση δασκάλου-μαθητή και την εκπαίδευση του. Οι είσοδοι του BERT και distilBERT, σε αυτή την περίπτωση, διοχετεύονται από τις βάσεις δεδομένων BookCorp και BookCorpus, όπου το δεύτερο αποτελεί μικρό υποσύνολο του πρώτου. Στη συνέχεια, το distilBERT εκπαιδεύεται εκ νέου, έχοντας αρχικοποιήσει τα βάρη του, με τα k-σημαντικότερα layers του δασκάλου. Η τελευταία διαδικασία είναι αυτή που ονομάζουμε μεταφορά γνώσης.



Εικόνα 2: Σχέση δασκάλου-μαθητή

ELECTRA

Ένα άλλο PLM που έχει επιλεχθεί για την εφαρμογή distillation είναι το ELECTRA, το οποίο είναι προ-εκπαιδευμένο . Το συγκεκριμένο μοντέλο αποτελείται από έναν generator και ένα discriminator. Η λειτουργία του generator είναι ίδια με αυτή που εξηγήθηκε πιο πάνω για το BERT. Το discriminator κομμάτι της αρχιτεκτονικής του, δέχεται ως είσοδο την έξοδο του generator, έπειτα γίνεται έλεγχος των predicted λέξεων. Ο έλεγχος γίνεται με κριτήριο, αν η λέξη που προβλέφθηκε από τον generator είναι η ίδια με την αρχική της εισόδου. Η έξοδος του discriminator είναι «original» όταν το κριτήριο ικανοποιείται αλλιώς είναι «replaced» . Η εικόνα 3 δίνει ένα παράδειγμα της αρχιτεκτονικής και του τρόπου λειτουργίας του ELECTRA.



Εικόνα 3: Αρχιτεκτονική ELECTRA

Ζητήθηκε να εφαρμοστεί το knowledge distillation(KD) στο ELECTRA, με τον ίδιο τρόπο που εφαρμόστηκε στο distilBERT. Μετά από βιβλιογραφική έρευνα, προτείνεται ένας τρόπος distillation για το ELECTRA με την παρακάτω διαδικασία. Επιλέγεται μόνο ο discriminator για το distillation του μοντέλου, ενώ το generator συστήνεται να έχει μικρό αριθμό layers. Έχει δοκιμαστεί να εφαρμοστεί knowledge distillation και στο generator χωρίς όμως σημαντική βελτίωση της επίδοσης του μοντέλου. Κατά τη μεταφορά γνώσης από τον δάσκαλο στο μαθητή εφαρμόζονται συναρτήσεις σφάλματος, έτσι ώστε να επιλεχθούν τα σημαντικότερα layers, με αποτέλεσμα να μεταφερθεί όσο το δυνατόν καλύτερη γνώση στον μαθητή. Συγκεκριμένα, η συνολική συνάρτηση σφάλματος ορίζεται σύμφωνα με τη βιβλιογραφία[5]:

$$Lkd = Lce + b1\ Lmse + b2\ Lcos + b3\ Lkl$$

Οι όροι αυτοί αναφέρονται σε:

b1, b2, b3: συντελεστές βαρύτητας

Lce: cross-entropy loss function to the student's classification outputs

Lcos: cosine loss function between student's and teacher's output activations

L***kl:*** KL divergence loss function with softmax temperature t student's and teacher's output activations

Lmse = Lemb + Lout + Lhid + Latt

L***emb*** = MSE loss to the input embeddings between teacher's and student's

Latt = MSE loss to the output of the self- attention heads between teacher's and student's

Lhid = MSE loss to the hidden layers between teacher's and student's

Lout = MSE loss to the output activations between teacher's and student's

ΕΚΠΑΙΔΕΥΣΗ

Δεδομένου ότι η εκπαίδευση έπρεπε να γίνει σε κάποια υπηρεσία cloud, ο χρόνος εκπαίδευσης ανά εποχή πρέπει να περιοριστεί σημαντικά, όπως και το σύνολο των δεδομένων εκπαίδευσης, λόγω των περιορισμών που υπάρχουν στο χρόνο χρήσης της υπηρεσίας και στη διαθέσιμη μνήμη που παρέχεται. Σημειώνεται ότι για την εκπαίδευση αξιοποιήθηκε ο κώδικας [2] που συνοδεύει το [1].

Ως σύνολο εκπαίδευσης χρησιμοποιήθηκε ένα υποσύνολο του dataset wikitext [3]. Η χρήση υποσυνόλου κρίθηκε αναγκαία, γιατί το μέγεθος είναι καθοριστικό τόσο για το χρόνο εκτέλεσης όσο και για τη χρήση μνήμης.

Ακόμη, το μοντέλο προς εκπαίδευση που χρησιμοποιήθηκε στην παρούσα εργασία έχει τη δομή του DistilBERT[1], με τη διαφορά ότι το πλήθος των layers είναι 4 αντί για 6. Ως teacher χρησιμοποιήθηκε το BERT(bert-base-uncased). Η επιλογή αυτή έγινε λόγω των περιορισμένων πόρων, αν και σημειώνεται ότι η μείωση του αριθμού των layers από 6 σε 4, δεν έχει τόσο σημαντική επίδραση στον χρόνο που χρειάζεται για την εκπαίδευση συγκριτικά με τις λοιπές τροποποιήσεις που έγιναν. Ενδεικτικά, η εκτέλεση για τα 6 layers στο υποσύνολο δεδομένων που χρησιμοποιήθηκε χρειαζόταν 180 λεπτά περίπου, ενώ για τα 4 layers 150 λεπτά (οι συγκεκριμένοι χρόνοι αφορούν training χωρίς τα optimization που αναφέρονται παρακάτω).

Για περαιτέρω μείωση των απαιτήσεων για την εκπαίδευση του μοντέλου, κρίθηκε χρήσιμη η χρήση floating points με ακρίβεια 16 bit έναντι 32 bit. Η χρήση μικρότερης ακρίβειας παρατηρήθηκε οτι μειώνει σημαντικά τον χρόνο εκτέλεσης. Σχετικά με τα optimization του compiler (flag fp16_opt_level) χρησιμοποιήθηκε η τιμή O2 (FP16 training with FP32 batchnorm and FP32 master weights), καθώς για εκτέλεση με flag O3 παρατηρήθηκε ότι το loss παραμένει σε υψηλές τιμές, οπότε δεν γίνεται training.

Τελικά, στο πλαίσιο της εργασίας εκπαιδεύτηκαν 3 μοντέλα για 3 εποχές. Τα 2 πρώτα αφορούν εκπαίδευση με τη κοινή συνάρτηση loss που χρησιμοποιείται στο [1]. Συγκεκριμένα, η συνάρτηση που χρησιμοποιήθηκε είναι η $L = 5\ L_{CE} + 2\ L_{MLM} + L_{COS}$. Το πρώτο αρχικοποιήθηκε με βάση τα layers 1, 5, 8, 12 του BERT, ενώ το δεύτερο με τα layers 1, 3, 4, 6 του DistilBERT. Εκπαιδεύτηκε επίσης ένα μοντέλο με αρχικοποίηση βάσει των ίδιων layers του BERT αντικαθιστώντας το cosine loss με ένα MSE loss μεταξύ των αντίστοιχων layers του teacher (BERT) και του student, δηλαδή χρησιμοποιήθηκε η συνάρτηση $L = 5\ L_{CE} + 2\ L_{MLM} + L_{ATT.}$

Στο [6] υπάρχει ο κώδικας που αξιοποιήθηκε στα πλαίσια της εργασίας. Στο README περιγράφονται ποιες αλλαγές έχουν γίνει, ενώ παρατίθενται και σύνδεσμοι για notebooks που δείχνουν πως μπορεί να γίνει η εκπαίδευση αλλά και η αξιολόγηση των μοντέλων με τον κώδικα αυτό.

ΑΠΟΤΕΛΕΣΜΑΤΑ

Αξιολογήσαμε την επίδοση των τριών μοντέλων που εκπαιδεύσαμε στο GLUE benchmark. Για την αξιολόγηση έγινε fine-tuning για 3 εποχές εξετάζοντας τις τιμές {1e-5, 5e-5, 1e-4} για το learning rate και τις τιμές {16, 32} για το batch size, εκτός από τα tasks QQP και MNLI

που απαιτούν περισσότερο από 90 λεπτά για την εκπαίδευση και την αξιολόγηση. Τα αποτελέσματα τους δεν είναι βέλτιστα, όμως οι τιμές είναι κοντά στις τιμές του DistilBERT, επομένως δεν αναμένονται σημαντικά καλύτερα για διαφορετικές παραμέτρους.

Τα αποτελέσματα απεικονίζονται στον παρακάτω πίνακα, όπου στην πρώτη στήλη αναφέρεται το μοντέλο από το οποίο έγινε η αρχικοποίηση (αν υπάρχει παρένθεση υποδεικνύει το loss function που χρησιμοποιήθηκε):

Μοντέλο	**CoLA**	**MNLI**		**MRPC**		**QNLI**	**QQP**		**RTE**	**SST-2**	**STS-B**		**WNLI**
	Matthews corr	Matched Accuracy	Mismatched Accuracy	F1	Accu	Accu	Accu	F1	Accu	Accu	Pearson cor	Spearman corr	Accu
DistilBERT	37.33	78.06	78.29	87.52	81.37	85.74	89.41	85.89	57.04	89.45	85.24	84.85	56.84
BERT	28.88	78.14	79.05	86.28	79.66	85.61	89.34	85.86	58.48	88.88	83.53	83.34	56.34
BERT (Att-MSE)	29.62	78.02	78.26	86.14	79.41	85.1	89.35	85.92	58.48	89.56	83.44	83.12	50.7

Πίνακας 1: evaluation 3 μοντέλων μέσω GLUE BENCHMARK

Παραθέτουμε έναν πίνακα για τα αντίστοιχα αποτελέσματα για το BERT και το DistilBERT, όπως προκύπτουν από το [1]. Σημειώνεται ότι για τα tasks που έχουν περισσότερες από μία μετρικές στον παρακάτω πίνακα δίνεται το combined score που είναι ένα macro average των επιμέρους μετρικών.

Μοντέλο	**CoLA**	**MNLI**	**MRPC**	**QNLI**	**QQP**	**RTE**	**SST-2**	**STS-B**	**WNLI**
BERT	56.3	86.7	88.6	91.8	89.6	69.3	92.7	89.0	53.5
DistilBERT	51.3	82.2	87.5	89.2	88.5	59.9	91.3	86.9	56.3

Πίνακας 2: Σύγκριση BERT και DistilBERT [1]

Αρχικά σε μία σύγκριση μεταξύ των μοντέλων που εκπαιδεύτηκαν για την παρούσα εργασία και των προεκπαιδευμένων μοντέλων, παρατηρείται ότι τα μοντέλα της εργασίας δίνουν αποτελέσματα συγκρίσιμα με αυτά των γνωστών μοντέλων. Εξαιρώντας το CoLA task για το οποίο όλα τα μοντέλα δίνουν σημαντικά χαμηλότερα αποτελέσματα έναντι των γνωστών, για τα υπόλοιπα τα αποτελέσματα η διαφορά δεν ξεπερνά το 10 % για κάποιο task. Η μείωση της επίδοσης είναι αναμενόμενη, καθώς ένα μοντέλο που έχει μικρότερο μέγεθος και έχει εκπαιδευθεί σε λίγα δεδομένα αναμένεται να έχει χαμηλότερη ακρίβεια. Παρ' όλα αυτά η μείωση δεν είναι ιδιαίτερα σημαντική καθιστώντας τα μοντέλα χρήσιμα για εκτέλεση σε περιβάλλον με περιορισμένους πόρους. Σημειώνεται, ακόμη, ότι για το task WNLI δύο από τα μοντέλα που εκπαιδεύτηκαν δίνουν καλύτερα αποτελέσματα από τα προεκπαιδευμένα.

Συγκρίνοντας μεταξύ τους τα μοντέλα, επιβεβαιώνονται τα αποτελέσματα του [4] ότι η αρχικοποίηση δεν είναι ιδιαίτερα σημαντική για task-agnostic distillation, καθώς τα δύο πρώτα μοντέλα, τα οποία διαφέρουν μόνο στην αρχικοποίηση, δίνουν παρόμοια αποτελέσματα. Συγκεκριμένα, το μοντέλο που έχει αρχικοποιηθεί από το DistilBERT δίνει στα περισσότερα tasks ελαφρώς καλύτερα αποτελέσματα, σημαντικά καλύτερα μόνο για το CoLA, το οποίο εξηγείται από τη μεγάλη ομοιότητα του μοντέλου που εκπαιδεύτηκε και του DistilBERT από το οποίο

αρχικοποιήθηκε. Όμως, η διαφορά δεν είναι ιδιαίτερα σημαντική και το training ήταν περιορισμένο για να θεωρηθεί ως σημαντικό πλεονέκτημα.

Συγκρίνοντας τα δύο τελευταία μοντέλο τα οποία διαφέρουν μόνο στο loss που χρησιμοποιήθηκε κατά την εκπαίδευση, παρατηρείται σημαντική διαφορά μόνο για την περίπτωση του WNLI, ενώ για όλα τα υπόλοιπα τα αποτελέσματα είναι παρόμοια, υποδεικνύοντας ότι δεν υπερτερεί η μία συνάρτηση loss έναντι της άλλης.

ΣΥΖΗΤΗΣΗ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ

Λόγω της αρχιτεκτονικής του ELECTRA που εξηγήθηκε πιο πάνω γίνεται αντιληπτό ότι η μεταφορά γνώσης είναι απαιτητική. Η παρουσία του discriminator όμως στην αρχιτεκτονική του ELECTRA καθιστά χρονοβόρα τη διαδικασία λαμβάνοντας υπόψη τους περιορισμένους υπολογιστικούς πόρους και το χρονικό διάστημα για την ολοκλήρωση της εργασίας.

Στα πλαίσια της εργασίας εκπαιδεύτηκαν 3 μοντέλα με τη δομή του DistilBERT, αλλά μικρότερο αριθμό layers. Παρά το μικρό μέγεθος και το σχετικά λίγο χρόνο για training η αξιολόγηση δείχνει ότι μέσω του Knowledge Distillation, μπορούν να εκπαιδευτούν αποδοτικά μοντέλα μικρότερου μεγέθους, τα οποία θα έχουν σημαντικό ποσοστό της γνώσης των αρχικών μοντέλων.

Σχετικά με την αρχικοποίηση, τα αποτελέσματα επιβεβαιώνουν τα αποτελέσματα του [4] ότι δεν είναι ιδιαίτερα σημαντική η αρχικοποίηση που θα επιλεχθεί για task-agnostic distillation για το σύνηθες loss που χρησιμοποιείται για distillation.

Τέλος, όπως προκύπτει από το [1] η χρήση cosine embedding loss μεταξύ των hidden states teacher και student είναι ωφέλιμη για το training. Από τα αποτελέσματα προκύπτει ότι η χρήση του MSE loss μεταξύ συγκεκριμένων attention layers του teacher και των attention layers του student, όπως περιγράφεται στο [4], αντί του cosine embedding loss δίνει παρόμοια αποτελέσματα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

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

[2]https://github.com/huggingface/transformers/tree/main/examples/research_projects/distillation

[3]https://huggingface.co/datasets/wikitext

[4] Wang, X., Weissweiler, L., Schütze, H., & Plank, B. (2023). How to distill your Bert: An empirical study on the impact of weight initialisation and distillation objectives. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. https://doi.org/10.18653/v1/2023.acl-short.157

[5] Hentschel, M., Tsunoo, E., & Okuda, T. (2021). Making punctuation restoration robust and fast with multi-task learning and knowledge distillation. *ICASSP 2021 - 2021 IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/icassp39728.2021.9414518

[6] https://github.com/ggotz/Distillation

[7] Clark, K., Luong, M., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. *arXiv (Cornell University)*. https://arxiv.org/pdf/2003.10555.pdf