

Transfer learning: Zengin ve büyük bir dataset

üzerinde pre-train ~~kip~~ edilen bir modelin daha sonra downstream task üzerinde fine-tune edilmesi

T5'in omzu bir modeli birden fazla

is için kullanılcak şekilde eğitme ve bilin işlenmesi

(Text-to-Text formatında işlem)

Model eğitim veri seti

Common crawl'dan alınan veribit

tilizlile işlenmiştir <sup>daha</sup> sonra 750 GB'lik CLU veri seti kullanıldı.

Model mimarisi:

Model mimarisi Vanilla transformer'a benziyor

ama ufak farklılıklar var

#### • encoder'da

input sequence'ları encoder'a iletmek için

Embedding katmanı - Encoder self-attention

layer ve FFN içerir. cyrlico Layer normalization

icerir ama daha basit bir versiyon. Layer Norm'da 3.1.1 Model

Sadece ölçütlerme yapılıp biraz uygulanmaya model olarak standart encoder-decoder bir Layer Norm kullanılır. Dropout FFN'de, Transformer modeli öneriliyor

#### • Decoder

encoder'e benzer bir yapı gösterir ama ~~self~~ farklı olarak Self-attention'dan sonra Klasik Attention mekanizması içerir. Decoder'da

Self-attention mekanizması, modelin yedinci gecmiş bilgileri görebilmesi için bir tane oto regresif layer ~~self~~ ~~self~~ causal self attention kullanılır.

Son decoder bloğunu ~~elde etmek~~ elde etmek, öğrenlikleri giriş gümme matrisi ile paylaşılan softmax çıktıları bir deneysel layer'la beslenir

Transformers'deki tüm attention mekanizmları, deneysel çoklu işlenmesi önce çıktıları birleştirerek birleştirilir.

Relative positional embedding kullanılır

cyrlico vanille transformer'da ~~512~~ farklı olarak layerının pozisyonu kullanılır

Downstream task:

~~other than the~~

translation

summarization

Performs benchmark

WMT

CNN/Daily Mail

Question answering

SQuAD

Text classification

GLUE and SuperGLUE

~~Input and Output Format~~

Yukarıda açıklanan çeşitli görevler üzerinde test bindeki

modeli eğitmek için, ele aldığımız tüm görevlerin "text-to-text" formatına çevrildi. Yani Modelin ~~bağlantı~~ text ile başleyip text üretmesi sağlanır

#### 3.1 Baseline ~~task~~

One model olarak standart bir transformer

Modelin: basit bir denoising hedefi kullanılarak

ON eğitine tabi tutularak diğer downstream

tasklar için fine-tune edilir

#### 3.1.1 Model

Model olarak standart encoder-decoder

skip connection'da attention weight'lerde,

ve tüm yığının giriş ve çıkışında uygulanır

• n-layer = 12 • n-head = 12

• d\_ff = 3072 • d\_kv = 64

• d-model = 768 • dropout = 0.1

• max-len = 512 • batch-size = 128

• lr = 0.01 • 220M params

• vocabsize = 32,000

#### Vocabulary

Metni WordPiece belirteçleri olarak ~~bir~~ kodlamak için Sentence Piece kullanılır

91

## Unsupervised Objective

Model: önceden eğitme için etiketsiz verilerden yorumlamak, etiket gerektirmeyen en çok modele sonraki örenmeye yararlı olacak görelleştirebilin bir deneme. Daha sonra bir alternatif bilgiyi öğretir bir hedef öğretti.

Modelin tüm parametreleri önceden eğitme ve ince ayar yapma transfer öğrenme paradigmasını NLP problemlerine uygulayan öncedeki dil modelleri ön eğitim için nedenel bir dil modelleri (coupled language modelling) hedefi kullanmıştır.

Ancak son zamanda (paper 19 Sep 2023'e eft)

"denoising" hedeflerinin (mosteklemiş dil modellerine göre de calanları) daha iyi performans gösterdiği gösterilmis ve hızla standart haline gelmiştir. Bir denoising hedefinde model, girdideki eksik veya boşluğunu belirteleri tespit etmek için eğitir.

Bartın "masked language modeling" ve "word dropout" denetleme tekniklerinden esinlenerek, restgele örenmeye yorum ve ordandan gittiğinizdeki ~~text~~ belirteğini ~~text~~'ini birkaç bir hedef toplandı. Her sentinel token, dizide özgü bir token kimliği etenir. Ardından belirteğin olumlu olmaması maskelene ve yalnızca ~~text~~ dropout edilen belirtekeri tespit etme şansları, ön eğitim similasyonlarını etkilemek için kullanılmıştır.

## Baseline Performance

İdeel olarak sonucular izinde given erdiği gibi etmeden için eğitim modeli her doreyi deneme. Daha sonra bir alternatif deneme baseline modeli 10 kez eğitiliyor (yani farklı restgele başlatmalar ve varsa seti karıştırma ile).

sonuçlar

	GLUE	CNERP	SQuAD	SWEB	Ende	EnFr	EnRo
Base line average	43.24	49.24	60.69	71.36	26.98	34.82	22.65
Baseline standard deviation	0.239	0.065	0.343	0.416	0.112	0.090	0.104
No pre-training	46.22	47.60	50.31	53.04	25.86	39.72	24.04

## original text

Thank you ~~for inviting~~ me to your party last week

## inputs

Thank you ~~(x>~~ me to your party ~~(y>~~ week

## Targets

~~(x>~~ for inviting ~~(y>~~ last ~~(z>~~

## 3.2 Architectures

Transformer ilk olarak bir encoder-decoder mimarisini ile tanıtılmış olsada, NLP için transfer learning üzerinde yapılan birçok modern çalışma alternatif mimariler kullanmaktadır.

### 3.2.4 Model Structures

Farklı mimariler için önemli ayırt edici faktörler, modeldeki farklı direkt (attention) mekanizmalarını tarafından kullanılan "maskleme". Bir Transformer'da self-attention işlemiñ girdi olarak bir dizisi olurken ve aynı ~~de~~ uzunlukta yeni bir dizisi oluşturduğu biliriz. Bu dizisinin her bir girdisi, girdi dizisinin girdilerinin oluşturduğu ortalaması hesaplanır.

~~ve aynı ~~de~~ uzunlukta yeni bir dizisi oluşturduğu biliriz. Bu dizisinin her bir girdisi, girdi dizisinin girdilerinin oluşturduğu ortalaması hesaplanır.~~

• Full-visible attention mode : Standard encoder-decoder Transformer mimarisinde encoder'da kullanılır. full-visible attention mode, bir self-attention mekanizmasının çıktılarının her bir girdisini üretirken girdinin herhangi bir girdisine katılmasına izin verir.

~~her çıktıda her bir self-attention mekanizmasının girdinin tüm ögesini kullanma engelmesi için var olan bosluk kullanılabilmektedir.~~

• Causal masking : Transformer'in decoderi

Causal mask kullanılarak maskeleme yöntemi

çıktı dizisinin ~~in~~ ögesi üretildikten modelin

gelecekteki ( $J > L$ ) giriş ögelerine bakmasını engeller

Böylece model eğitimi sırasında geleceği görmeden sonradan next token ile tahmin yapar

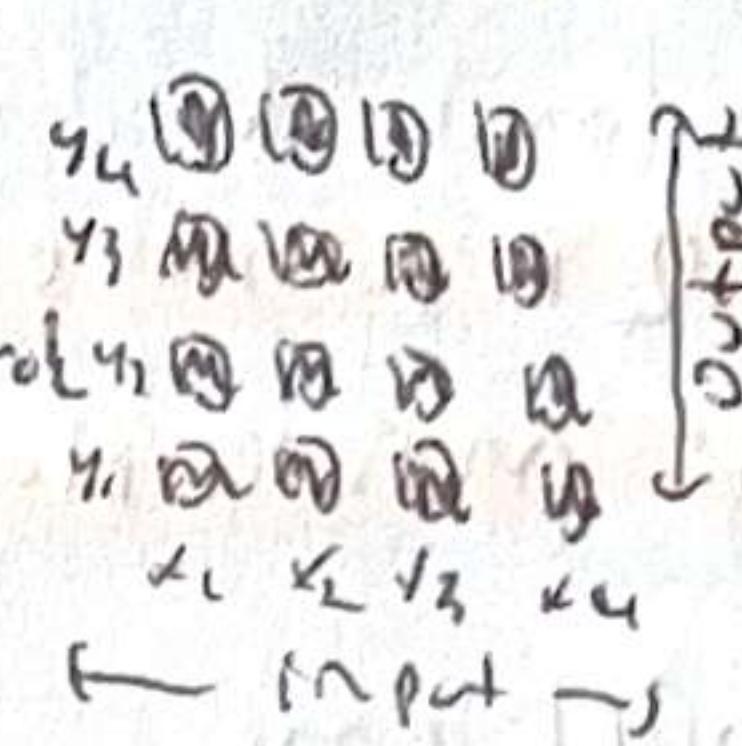
Dil modelleri genellikle sıktırma veya sequence душтрумак için kullanılır. Ancak, girdileri ve hedefleri birleştirerek text-to-text framework'inde kullanılabilir - ingilizceden Almancaye çeviri için

Modelde "translate English to German = that is good-target": ön ekler verdir ve dizinin geri kolunu, otoregresif olarak üretmesi istenir

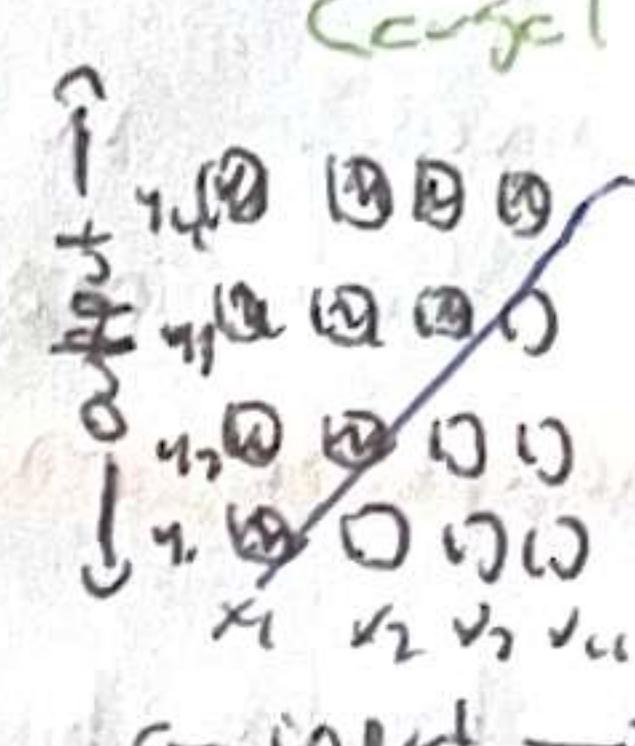
text-to-text également için bir LM kullanmak desavantajlarından birisi causal masking'in in sedice i. bölümde keden dan gideye direkt olmasi bunu aşmak için causal with prefix masking ortaya çıktı

• Causal with prefix : Dil modelinde gelenekselde ek olarak belirlili bir seqit başlangıcı (prefix) kullanılarak buna izin veren bir maskeleme yöntemidir

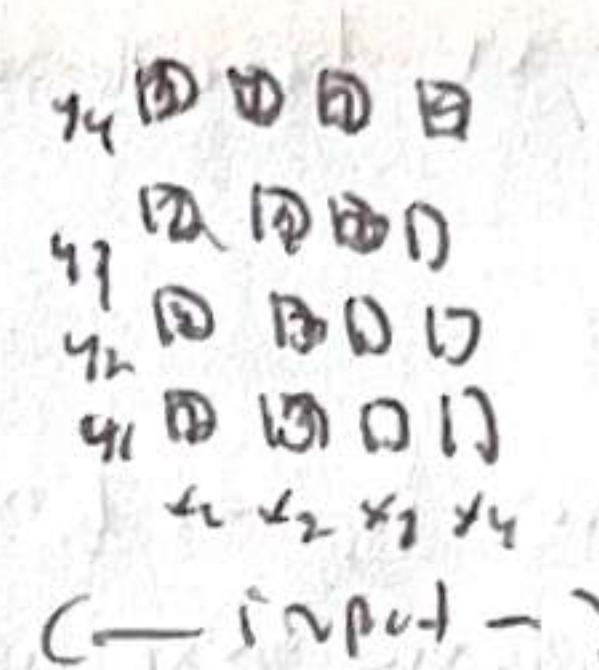
full-visible



Causal

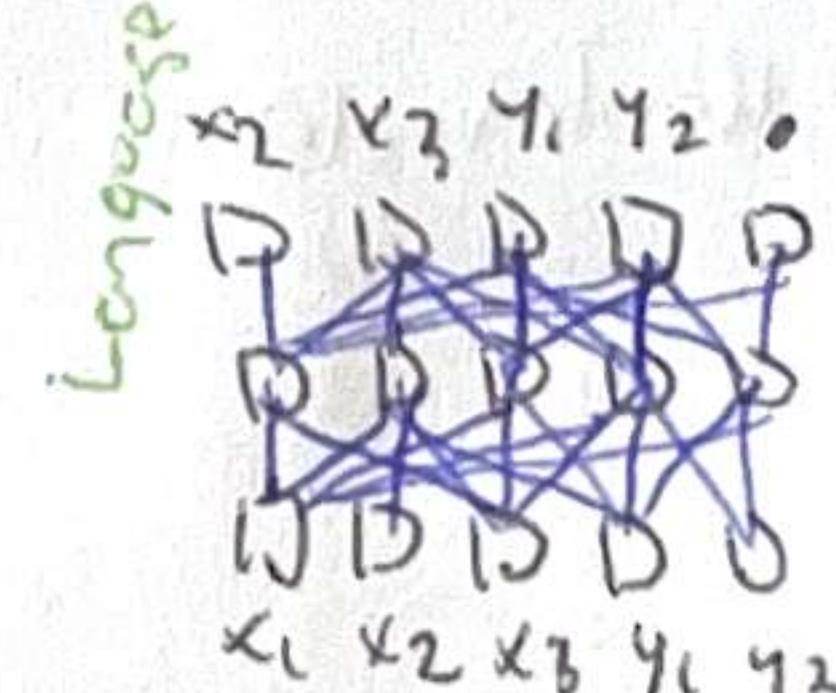


causal with prefix

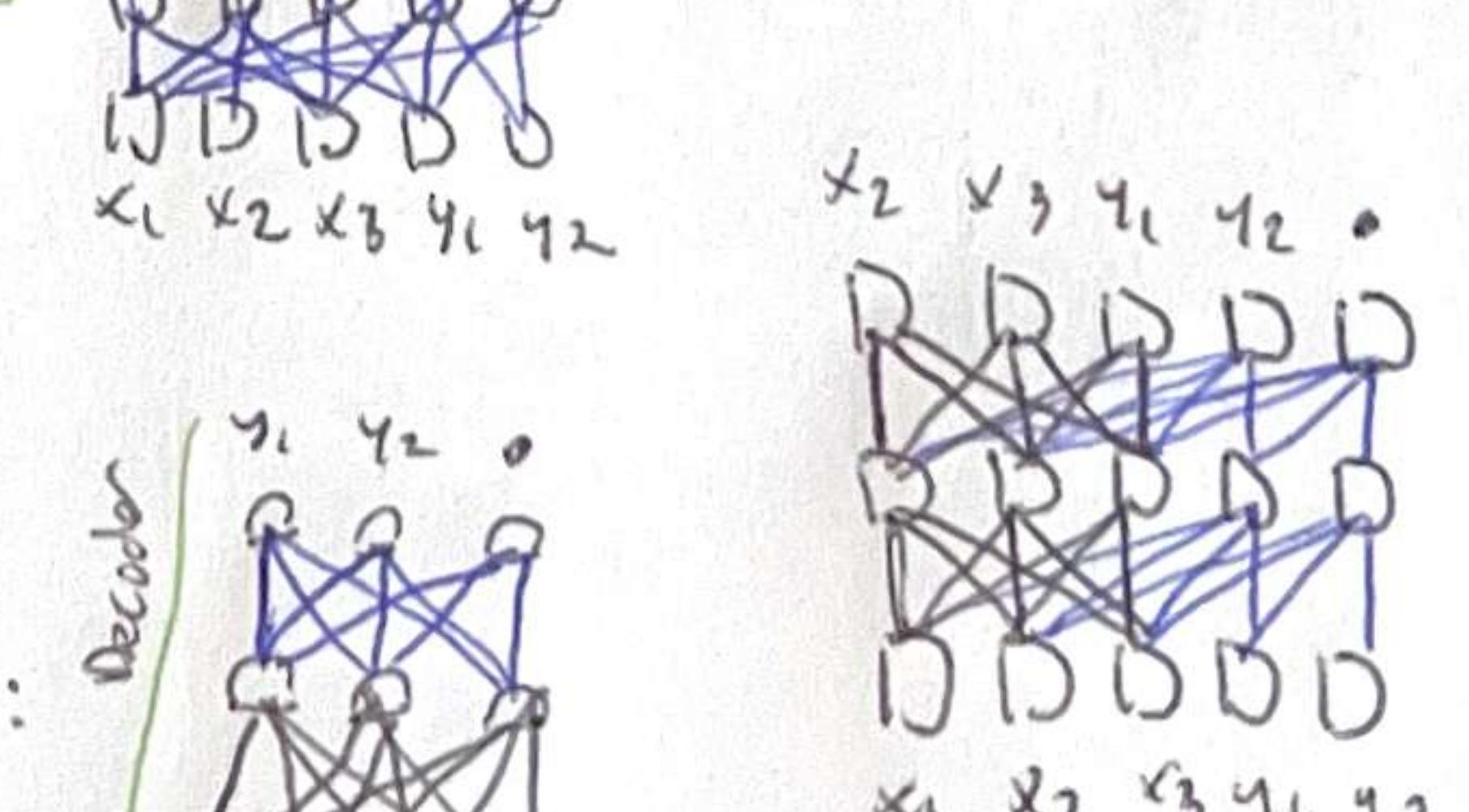


→ full-visible attention

→ causal masking



Prefix LM



### 3.2.2 Comparing Different Model Structures

L katmanlı schip bir encoder-decoder modeli  
ile 2L katmanlı schip dil modeli ile aynı  
sayıda parametreye schip tira öteyordan

L+L encoder decoder modeli L katmanlı  
dil modeli ile aynı hesaplama maliyetine schip'tir  
Bu faktur sebebi dil modelindeki L katmanının  
herhangi giriş hende此文 dizisine uygunlaşmasından,  
encoder'in sadece giriş dizisine ve decoder'in  
yolnırca此文 (target) dizisine uygunlaşmasında  
kaynaklanmaktadır

- L katmanlı encoder ve L katmanlı decoder  
iceren encoder decoder modeli **2P parametre  
icerdir ve hesaplama maliyeti M FLOPs olur**

- Yukarıdakine benze bir model cme  
encoder ve decoder arasında parametre paylaşımı  
olan bir model **P parametre içeri ve hesaplama  
maliyeti M FLOPs olur**

- Encoder ve decoderde **L/2 katman içeri bir**  
model **P parametre içeri ve hesaplama maliyeti**  
**M/2 FLOPs olur**

- ~~Decoder-only~~ bir dil modeli L katmanlı  
ve P parametre içeriysse **hesaplama maliyeti**  
**M FLOPs olur**

- Decoder-only prefix LM normal decoder-only  
LM ile aynıdır

- Transformers blokları arasında parametre paylaşımı  
performansın öden vermeden toplam parametre  
sayısını azaltman ettili bir yolu bulmuştur

- Bir denoising hefdefini kullanmasının  
bir dil

Unsupervised next word  
uygulandıca genel amaci bilgi edindiği  
mekanizması sağladılar bir öneme schip'tir

ış takımları giriş  
odel (diziler)  
olo gerekliydi  
oyun performansı  
cripted token  
rulespace  
- some wrongin  
BERT-style  
önce)

en basitse  
vitepi:

i. Sınırlar

değişkenligi

genel özellikle  
or

dünyede  
ortalaması  
(10)

fark yok

sunelamış  
məcləməsi  
lətin  
optimizasiya

Inputs	Targets
obtree live	that you for inviting
prefix language modeling	that you and invite me to your party opple week
BERT-style Denman et al (2018)	party me for you to let fun you inviting week than
Deshoteling	party me for you and invite me to your party lns week
MASS-style Sung et al (2019)	that you lns and me to your party lgs week
Tid. noise replace song	that you me to your party lgs work
Tid. noise drop tokens	lgs for inviting lgs last (22)
Random song	for inviting losi
	(+) for inviting me lgs your party last(22)

Q

matematik birlegimizde

5

Redore珊瑚礁

### 3.3.1 DI sporcle High-level Approaches

İlk olcuk yoğun olacak kullanicilar 3 farklı yoldan  
kayboluyor - İlk olcuk PrefixLM hedefi.

↳ Bir tek nitelik bir metin parçasını bir.

↳ encoder'a girdi olarak kullanicilar, digeri ise  
decoder toknular teknisi edilecek bir hedef dizisi  
olarak kullanacak iken bilesen ayırmak 2. olcuk  
BERT'le kullanicilar "Masked language modelling" (MLM)  
hedefinde esitlenerek bir hedefi ele aliyoruz. MLM bir  
metin analizi olur ve belirteklarin 90%'ini bozar.

Bozular belirteklarin 90%'i özel bir maske belirteci  
ile degistiriliyor ve %10'u restgeli bir belirteci ile  
degistiriliyor. BERT yarizco encoder-only bir model  
oldugundan, pre-train sirosndeki amaci encoderin  
silgisindeki maskelemis belirteklar yerida yapildirmaktır.

Encoder-decoder durumunda, hedef olarak bozulmasın  
dizisinin tamamen kullanilmasa - Son olcuk bir de  
dendensing sequential autoencoder'yu uygulanan temel  
bir deshuffling hedefi ele alınır. Bu yoldan bir  
belirteci dizisi olur, kariyfim ve orijinal  
kariyfim gibi giderilmis dizisi hedef olarak kullanır  
karilastirma soncu Guel olcuk, BERT terzi  
hedefin en iyi performansi gösterdiği gözlemlendi.

### 3.3.2 Simplifying the BERT Objective

Onceki bölümdeki sorulara dayanarak (Simdi BERT terzi  
terzi derosing hedefindeki degisibiltiler bekliyor).

Bu hedef baslangicta siniflandurma ve arclik teknigi  
icin egitilen encoder-only model icin bir pre-train  
tekniği olarak önerilmiştir - Bu nedenle, encoder-decoder  
text-to-text kurulumda dcho siyi performans göstericek,  
veo dcho usulü olacak şekilde degistirilmesi  
mungkin olabilir.

• MASS-style: Restgeli belirteci degistirme odiminin  
dehli olmadigi BERT terzi hedefin basit bir varyanti.  
Ortaya cikan hedef girdi belirteklarin 90%'ini bir  
maske ile degistirir ve model, orjinal bozulmasın  
dizisi yerida yapildurmak ekin egitiliyor.

• Varz yokntmler: Decoder'da cum dizileri üzerinde  
self-attention gerektiren cum bozulmasın metin  
arkiliklerini teknisi etmekte beginmeyi yolları

- a) : Baseline: Bozulmasın her token yine  
maske boyink yine, Ardisik bozulmus tokenlerin  
tamami benzeriz bir maske token ile degistiriliyor  
Hedef draf, her bire girdide yine gelen maske  
tokeni ile öneklemlenmiş "bozulmus" arkiliklerin bir legimindan  
olugut

→ Replace corrupted spans

- b) drop corrupted tokens: Bozulmus tokenler giriş  
dizisinden tamamen siliniliyor - Model digerler  
tokenleri sırayla yerden degistirmek suretiyle

Sonra olcuk hapsi hem hem aynı performans  
gostermizdir istenilen olur drop corrupted token  
yontemi GLUE öncelik 1.5'te bir iyileşme  
segiscedo SUPERGLUE'de dcho katı sırası vermistir

13. ite yoldan hem next nevedeye BERT-style  
ile aynı (bert replace corrupted dcho iki)  
hemde dcho öz maliyetli!

### 3.3.4 Corruption Spans

• Mevcut yoldan her girdi token 90% bozulma  
ördes degisimli (i-i-d) bir token writesi:  
bozulanın bozulmasının

• Ardisik birler fazla token bozulduğunda, buna  
bir "span" (colekt) olacak ele oluyor  
• Cum span tek bir benzeriz maske ile degistiriliyor

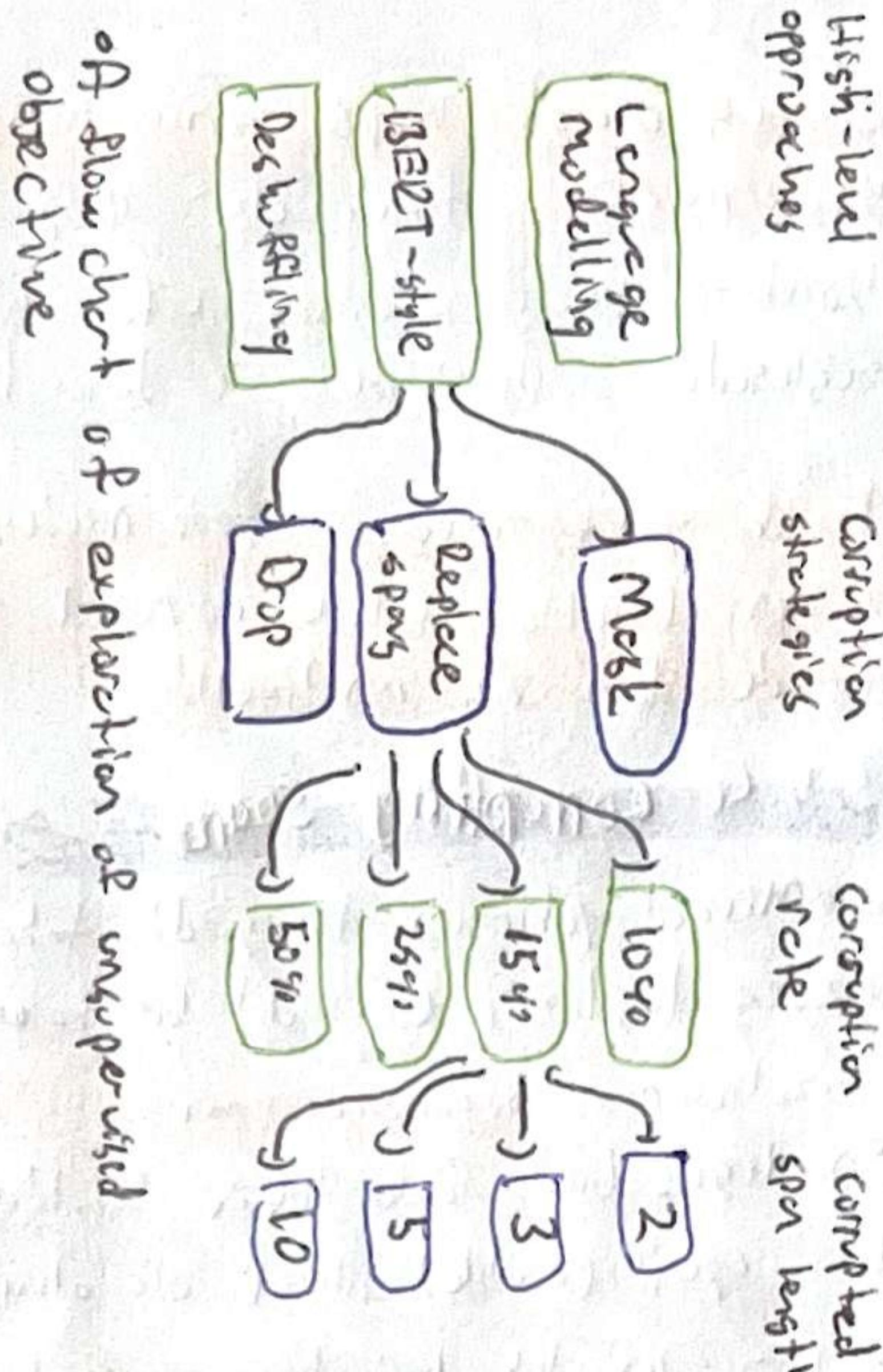
Arastirmalar basimsız tokenler bozulmak yerde özellikle  
bitizlik spans bozus öreme duruslar

bu bir dergi ile karilastirilmis - Cum dengelerde  
dls bozulma oranı kullanilmasa ve facta ortalamalı  
span ~~satılık~~ ozelliklerde donanmış (L, T, S ve so)

• Guel olcuk yoldasınlar orasında 50% fazla yok  
ordene span 10 olun bozun dcho katı sonelamis  
urdeme span 3 olun bozun dcho iyi sonelamis

Bu arastirma (3.3.4) büyük dil modellerinin  
egitim varmılığını doğrudan etkileyen bir optimizasyon  
strategisini inceliyor

### 3.3.5 Discussion



- Az miktarla token sorun değil, verilerdeki tokenlardığında ( $2^{24}$  token) etki sınırları 14 ya da miktarla token zararlı olmaya bilir.

- Mıktanın 13'te veri seti kullanılmıştır.
- Başka bir (daha büyük) model, daha büyük bir pre-train veri-setine overfitting yapmaya daha uygun olabilir.

#### 3.5 Training Strategy

~~standard token extractor~~

- Tüm parametrelerin üzerinde yapmakla berabere deki kaynaklı görevlendirme optimál olmaya sonucu yok olabilir.
- Üçüncü veri setini ile birlikte, modelin yapıları her parametresinin taraması gereklidir. Bu, overfitting'e neden olma riski var.

#### 3.5.1 Fine-Tuning Methods

Araştırmacılar yalnızca tanım parametrelerin her alt katmanını güncelliyor ve 2 katman üzerinde çalışıyor.

- Adaptor layers**: original model'in eğitimi sabit tutularak yalnızca kırıtkı katmanları eğitmeni easier.
- Transformer bloklardaki mevcut FFN'lerin her katmanı için "Dense - ReLU - Dense" blokları eklenir.
- 13+ katman eklenir, çıktı boyutu boyutlarının genelleştirilebilir şekilde tayin edilir. 135'yece modelin yapısında her katman değişiklik yapmadan eklebilir.

- Fine-tuning sırasında, yalnızca adaptor katmanları ve katman normalizasyon parametreleri güncellenebilir.

- Gradual UNfreezing**: zero içinde modelin her katmanının fine-tune edilmesi sağlanır.

- Fine-tune başlangıcunda, yalnızca son katman parametreleri güncellenebilir.

- Belirli sayıda güncelleme işlemi之后, sonraki 2 katman parametreleri da bil edilir. Böylece devam edilir.

- Tüm parametreleri fine-tune etmek genellikle en iyi performans verir ama hesaplama boyutunda maliyetlidir. Adaptor katman gibi kaynakları kullanırsak, daha hızlı bir alternatif sunabilir.

→ Bölüm 3.3.1'de farklı yoklamaları karşılaştırılmış ve BERT-style denoising hedefinin en iyi performansı gösterdiği belanmış.

→ Bölüm 3.3.2'de BERT yoklamasını değiştireerek daha büyük hedef dizileri değiştirmeyi denemisler. En iyi seçimin Replace spos teknigi olduğunu göstermiştir.

→ Bölüm 3.3.3'te farklı boyutlu ortamları (9, 10, 15, 25, 50) ile deneysel yapılmıştır.

→ Bölüm 3.3.4'te 5 farklı tokenin cardinalını katsayı olarak boyan bir hedefi deģerlendirmiştir.

Araştırmacılar, performansı korurken ve deyleştirirken eğitim sürecini daha verimli hale getirmeyi amaçlayan farklı yoklamaları ve parametrelerin adım adım degradasyonlarını

#### 3.4.2 Pre-training dataset size

Araştırmacılar modelin pre-train için kullanılan veri seti boyutu ve ~~verilen~~ verilen tokenin eğitimini nasıl etkilediğini en azından iki deneysel yapılmıştır.

→ Veri seti büyüdükçe (token sayısını artttır 40), performans genellikle düşüyor.

→ Araştırmacılar performans düşüşünün veri setinin erken kademelerde kaynaklı olabileceği能力和ını düşünenler

## Multi-task Pre-training

Bir modelin farklı görevler üzerinde aynı anda eğitilmesi sürecidir. Bu yöntem modelin genelleştirme yeteneğini artırmakla birlikte görevlerde daha iyi performans göstermesine yardımcı olabilir. <sup>1) Spesifik görevlerde yüksek performans istenirse</sup>

### 1 supervised multi-task pre-training

Model, farklı görevler için etikelli veri setleri ile eğitilir bu yöntemin avantajı:

- OMOS: Modelin çeşitli dil anlayışı görevlerinde performansını artırmak

örnek görevler:

- Sentiment analysis
- Question answering
- Natural Language Inference

örnek model:

MT-DNN (Liu et al., 2019b) <sup>Görel d.l. yeteneklerini kolaylaştırmak için</sup>

### 2 unsupervised Multi-task pre-training

Model self-supervised learning yöntemleriyle eğitilir

- OMOS: Modelin genel dil anlayışı konusunu sağlamada sonra her spesifik görev için özel optimize edilir.

- Masked Language Modeling (MLM) (BERT)

- Causal Language Modeling (CLM) (GPT)

(Sadece sezon belirleme teknini yapar)

- Next Sentence Prediction (NSP) (BERT, Cilti cümle ordutunu değiştirmi)

- Denoising Auto-encoding (T5, Gorüntü netağı taze hale getirir)

örnek model:

BERT, GPT, T5

## Combining Multi-task learning with fine-tuning

~~1. Modeli örnek sayısına uygun kılavuz kılavuz~~

~~2. Bu bölümde multi-task learning'in biraz daha detaylı ~~ve~~ ~~spesifik~~ versiyonunu gengeleterek~~

2. Modelin tüm görevler üzerinde pre-train edildikten sonra

3. Tek tek spesified görevler üzerinde fine-tuning yapılması,

önce okur bu yaklaşımın 3 varyantı var

- 1-) OMOS: Examples - proportion and mixture kılavuzları ile eğitilir.
  - Ancak bu rada veri seti büyükliğine  $E=2^{19}$  gibi bir sayıda veri koymak istenir. Sonrasında, her görev için ayrı ayrı finetune ediliyor.

- 2-) Supervised general unsupervised hybridlerle birlikte bir eğitime dahil edilmesinin, modele erken görevde faydalı bir şekilde MORUZ kılavuzunu sağlayıp sağlanmadığını ölçmek
  - Farklı kaynaklardan gelen farklı verileri karıştırarak, modelin daha genel bir beceri seti edinmesine yardımcı olmak istenir. Böylece eğitilmesi gereken görevdeki dikkat分散 (distribution) gözlemlenir.
  - Fine-tuning sırasında modelin belki bir görevde daha iyi adapt olup olmadığı test etmek

- 3-) Yoldaşının genel mantığı, önce genis scope eğitilmiş bir model oluşturup (daha önce okur) sonra her spesifik görev için özel optimize edilir.

- 2-) "Leave-One-Out" Multi-task training
  - İnceki varyantta tüm görevlerin birlikte bir eğitilmesi ve ardından her biri için ayrı ayrı fine-tuning yapılmıştı. Bu 2. varyantta OMOS modelin genel beceriler konusunda konu modlığını onlasmak için, multi-task pre-training'de her görevi bilinci derken diğer görevleri yoksayaştırır.

- Modeli yine examples - proportion and mixture ile bir eğitiriz ( $E=2^{19}$ )

- Ancak her iterasyonda bir görev (down stream task) bir eğitində sıklıkla çıkarılıyordu

- Sonrasında sıklıkla görev için ayrı bir fine-tuning uygulanıyor

- Bu 3. varyant tüm görevler için tek tek tekrarlanıyor

- Bu yöntem, gerçek dünya senaryolarını simüle etmek için faydalıdır çünkü önceden hiç görülmemiş bir görevin fine-tuning yapılıp yapılmadığı modelin adaptasyon yeteneği test edilmemiş olur

3-) Sadece supervised görevlerde pre-train  
fakir ne?

• İlk 2 yontemde her supervised ve unsupervised  
görevde on egitimde kullanılmış.

• Bu yontemde ikisinde supervised görevler cibartılıyor  
ve yalnızca supervised görevler ile on egitim  
el yapılıyor

### Sonuçlar:

• Multi-task pre-train + fine-tuning

• baseline (pre-train + fine-tuning) ile

• boyutlabilir performans sonuçları gözlemlenmiş

• Leave-one-out yontemi sadece biraz daha  
düşük performans gösterdi

• Bu multi-task pretraining'in yeni görevlere  
ciddi şekilde engellemesini ve modelin görevler  
arası aktarımı (transfer learning) yeteneğini  
gözle olduğu gözlemleniyor

• tamamen supervised görevlerde yapılan on egitim ensemble de belirsiz işlevse sağlıyor  
ve genel görevler hariç tüm görevlerde daha iyi  
performans sağlıyor

Model  
• en iyi yontem unsupervised pretraining + fine-tuning

• BBR Scaling

• Genel olarak model boyutunu artırmak ve egitim  
süresini uzatmak her zaman faydalıdır

• Modelin egitim scoresini uzatmak veya batch size'i  
artırmak benzer şekilde faydalıdır

• Ancak en iyi sonuçlar genellikle model boyutunu  
artırmayıyla elde ediliyor

• 313 m 2-) Egitim scoresi ve model boyutunun ~~artırması~~  
ortthamının etkileri birbirlerini toplayabiliyor  
2 kat büyük bir modelin 2 kat üzerinde

113 m esitmesi ile 4 kat büyük bir modelin standart egitim  
esitmesi ile çok daha hizigin bir faktör gösternmiştir

Modern d • Bu da boyutme ve daha uzun egitim scoresinin  
sistemde birbirini dengeleyen stratejiler olduğunu gösteriyor

kat mod

• Fine-tun

• Özellikle büyük ölçekte görevler için uyandır

3-) Batch size'i artırmak ve egitim scoresini  
artırmak benzer etkilere sahip olabilir

• batch size'i 4 kat artırmak, egitim 4 kat  
uzatmak kadar ettili

• Ancak, farklı öğrenme dinamikleri nedeniyle sonuçlar  
her zaman aynı olmaz (Shallue et al. 2014)

4-) Ensemble modeller belki görevlerde çok daha  
iyi sonuçlar veriyor

• CNR/DM, WMT EnGr, WMT EnRo gibi  
görevlerde 4 ayrı modelin ensemble, doğrudan  
ölçümleme yöntemlerden daha iyi performans göstermiştir

5-) Ensemble modellerin farklıları boyasızda

• Dört tane man cyr modelin birleştirilmesi,  
en iyisi sonuçları veriyor

• Tek bir modelin on egitiminden sonra farklı  
sekillerde ince ayar yapılışıyla diğer farklılıklar

• Dört tane man cyr modelin birleştirilmesi,  
en iyisi sonuçları veriyor

5-) Daha büyük modellerin ince ayar ve  
etkisi scoresi daha pehdidir

6-) Küçük bir modelin daha uzun egitim  
maliyet açısından daha verimli olabilir

• Eğer model birçok farklı görevde kullanılarak  
çoklu bir şekilde otomatik olarak scoresi egitilmiş  
bir modelin maliyeti daha iyi amortı  
edilebilir

7-) Ensemble teknikleri ile büyük model  
egitimini orasında bir maliyet düşmesi verdir

• 4 ayrı modelin egitimmesi tek bir büyük  
modelin egitimmesiyle benzer maliyetlidir

### 3.7 Putting It all Together

Objective : SPAN - corruption objective,

bütün SperBİERT'ler esinlenerek olğuları maske  
temel i̇id. denoising yöntemini yarına kullanmış  
(3- yonten ortalama 3) ve en fazla 3-ndan önce  
origində dizinin 4, 15; bəzən növbətiñ. sonra  
olcək bə yonten Həm birçə dəhə iyi perform  
göstərmis. Həmde heçfə dizinin dəhə kisə  
sayesinde hesaplama çağında birçə dəhə var  
dənmişdir.

- Longer Training = Och. von einem ersten

Performansla örenli: artis soglomistir (burada verigî çok fazla kelimeler etmek overfittinge neden olabilir). Temel model BERT'lerdeki XLNet'in 16 kat, RoBERT'a da 64 kat daha az pre-train dengesi, Ama CL detaylarının bâzılılığı, ~~verileri~~ verileri kılınmadan çok daha ~~faizli~~ düşük öğrenim yopilmesine olanak tanımıştır.

Model Size :

- Base Model : - TS'in Vorschila modeli (720 m param)
    - BERT (SAGE) boyutunda bir encoder-decoder kullanıyor
  - Small Model (60 m parametre) - Daha küçük ve hızlı
    - 6 encoder ve decoder katmanı, dcho diziçik model boyutu  
ve dcho öz parametre ile hesaplanır. 6 katmandan dcho boyutu  
değişmez.
  - Large Model (720 m parametre) : - BERT(LARGE) boyutunda  
bir encoder-decoder kullanıyor
    - Dcho derin bir og, dcho fazla fazla parametre ve geniş  
hesaplama gereklimidir. var
  - B13 model (2-5 B parametre) : - dPR (16,384)  
head-size = 32, dkv = 128
    - Dcho hizmet matrisi  
körperlerini için optimize edildi.
  - L13 model (11 B parametre) : dPR = 65,536
    - 128 - headed attention ile çok fazla hesaplama gücü gerekiyor
    - Modern donanımlarda (TPU gibi) en hızlı şekilde  
çalışmaktadır, özel olarak ölçüldü.

- Niche modeller:
  - Niche or nesoplano 90% lezim
  - Fine-tuning of 1%

- Fine-tuning ve inference için doldurulmalıdır

Büyük modeler: decho iyi performans sonuçları özellikle büyük ölçekte görevler için uygundur.

• Genelde büyük ölçekte görevler için uygundur

• Multi-task pre-training.

- Supervised ve unsupervised workerle  
Multi-task pre-training yapmak, sadece  
her unsupervised pre-train yapmak kadar etkili  
değil, Buna göre, fine-tuning işlemesinde kodor  
beklemek yerine, modelin genel performansını  
çoktan eğitmen sürecinde izleme faydet, sürüyor  
(most).
    - Bu yuzden overfitting yaşanmasın  
icin deha fazla etiketsiz veriyle eğitilmesi gerekir
  - Fine-tuning on individual GLUE and SuperGLUE
    - Tüm GLUE/SuperGLUE görevini bir arada  
fine-tune etmek, süreci yönetmeyi kolaylaştırır  
ancı görevlerde performans düşmesine sebep olabilir.  
Her görevi ayrı ayrı fine-tune etmek her  
görevde deha işi performans sc̄ılarken, deha  
ni içeren görevde overfitting sistem, arttırmak  
overfitting riskini azaltmaz.
    - batch size 16-32-64-128-256-512-1024-2048-4096  
- checkpoint d̄nci sıkılık artırıldı.

• Eger forbli görevler icin fine-tune  
yapiliyorsa, once lm görevler bolükte fine-tune  
edip performans degraderdir.

- Eger bir görevde dizi performansını iyorsa, o görevler için fine-tune edilebilir
- Dikkat: her iki görevde batch-size'ı kizaptıracak şekildecheckpoint alınır, ~~overfitting~~ overfitting önleme için önemlü bir strateji olabilir

child Beam Search ~~Beam Search~~

Sıktı dizilerin üretme şartının, den  
görevide greedy decoding'ın göre daha fazla

• Beam width 4, her cdurudo 4 Rurbu cdag  
~~dizisim~~ dizim degenerendirmesmi. soşler,  
bu segitliklilik ettirirken doğru ekti sageri etteri  
row 11

• Uzantı  $\approx$  0.61 (d = 0.61) tıllanır  
Modelin gereğinda uzun diziler istenir

• 13- Strateji, özellikle 'Güven' ve özetleme  
sibi görevinde performansı artırmak

- Test Sets: Modelin son değerlendirmesi, test setleri üzerinden yapılır.
- Geviri: İkinci WMT'ın yıllık newstest veri seti kullanıldı.
- Özetleme için CNN/Daily Mail test seti kullanıldı.
- GLUE ve SuperGLUE için resmi benchmark sonuçları tez edildi.
- SQuAD test seti için yetkilisi dinamik değiştiğinden validation sonuçları raporlamaya devam ediyor.
- Genel başarısı: T5 24 benchmark 18 inde State-of-the-art (SOTA) performansı sağladı.

En iyi performans 11 milyar parametreli model ile elde edildi.

Bu analiz T5'in başarısının büyük ölçüde öbeklenme ile ilgili olduğunu çok iyi göstermektedir. Özellikle önemli bir rol oynadığını gösteriyor. Özellikle büyük modellerin sadece fazla parametreleri değil direktör değil, deha iyi tasarruf kurularak da fazlalar faydaladığıergusun.

## 4. Reflection

### 4.1 Takeaways

- Text-to-Text: Tek bir model, iki görevi T5, bir NLP görevini: text-to-text framework içinde eder. Geçmişte text-to-text framework - aynı kayıp fonksiyonu ve kod çözme prosesinin kullanımını çok fazlı metin görevlerinde kullanarak, bu görevi gerçekleştirdi. Model, eğitmek için basit bir yol gösterdi. 13'süligine regmen text-to-text framework'ının görevi öznisi mimariler ile konsantre olabilir. performansı elde ettiği ve ölçülebilir bir şekilde birleştirildiğinde, nihai etki SOTA sonuçları ürettiği gözlenendi.

- Architectures: T5 modeli için farklı mimariler önerilmiştir. Ancak, text-to-text framework içinde en iyi performansı original transformer framework'de, encoder-decoder yapısı göstermiştir. Encoder-only (BERT gibi) veya decoder-only (GPT gibi) modellerde kaydedildi. encoder-decoder modeli iki kat daha fazla parametreleri sadece beraberhesaplama malzemesi gerektir. Ayrıca, encoder-decoder parametrelerinin paylaşılmamış parametre sayısını yarıya düşürmektedir. Performansa kayda değer bir düşüş yaşanmaktadır. Bu da, modelin verimli bir şekilde öğrenilebilmesini ve transfer eğitimi için yeni bir yapı sunmaktadır. (16)

- unsupervised objectives: restyle hizalma metrici yeriden yapılmadır. Deha "denoising" yokluğunun text-to-text framework'de beraber performans sergilediği gözleniyor. Bu nedenle deha bu hedefler arasında dizler içten hedeflenerek edilmesi öneriliyor. Bu sayede unsupervised pre-train deha hesapla ve verimli olur.
- Data Sets: CU veri seti kullanıldı.
- Pre-train sürecinde kütüphane tekniklerin veri kimlikleri performansı olumlu etkilemektedir.
- Training Strategies: Eğitim stratejileri, önceki ögesinden önce eğitilmiş bir modelin tüm parametrelerinin ince ayar sürecinde güncellenmesinin deha öz parametresi güncellemeye yönelik olarak yapılanmış yöntemlerde deha iyi sonuçlar verdi. Ancak tüm parametrelere güncellemek çok maliyetlidir. Multi-task learning'in en büyük zararı her görev için eğitim alanlarını belirtmektedir. Sonuç olarak güncellenen kırılgınlıkları aşmak için bulunan strateji, deneysel ince ayar sonrası gerekli olduğuları denetimsiz bir eğitimde karşılaştırıldığında deha iyi performans sağlanmadı. Ancak farklı görevlerin karışımıyla yapılan pre-train yapısında, yapılar ince ayarın deneysiz bir eğitimde beraber performans gösterdiği görüldü.

- Scaling: model deha farklı veri ile eğitmek, deha büyük bir model eğitmek ve modellerin bir ensemble'ini kullanarak performans üzerinde önemli bir soğlaklıği oluştururlar. deha büyük bir model deha farklı veri ile eğitmeli, genellikle deha büyük bir model deha öz ölçümde eğitmelidir. deha katı sonuçlar verir.

### • Putting It All Together: