# Рачунарска интелигенција

Велики језички модели – место и структура

Владимир Филиповић

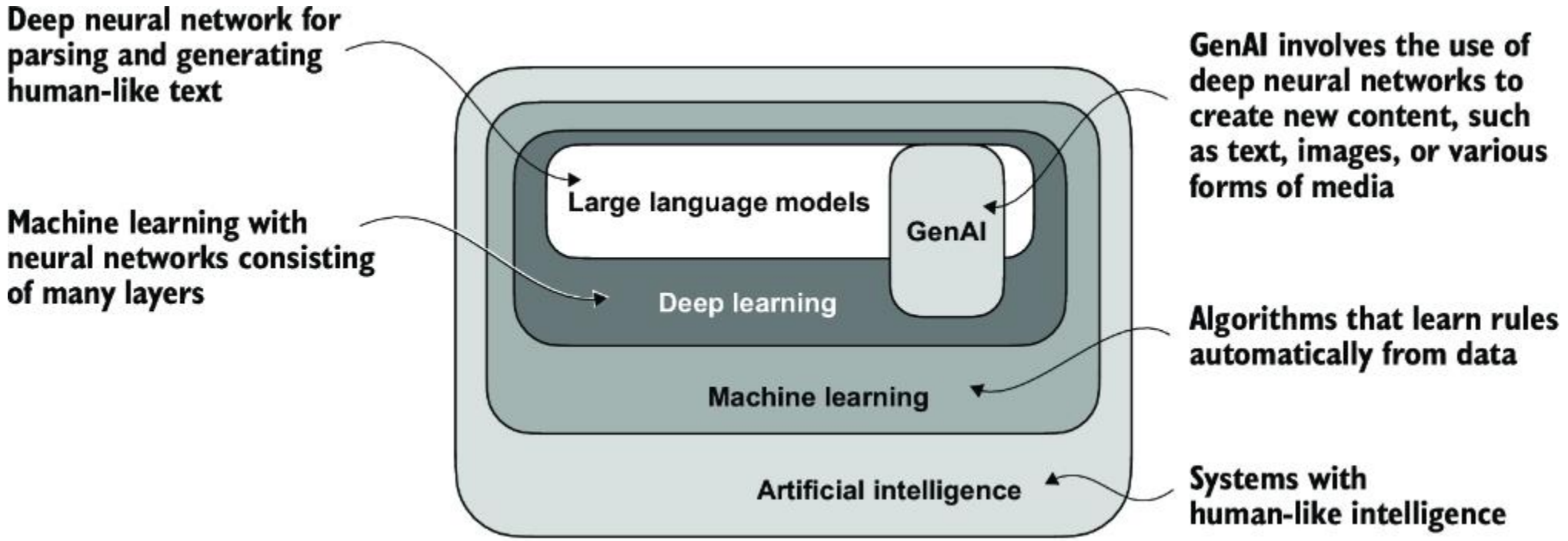
vladimir.filipovic@matf.bg.ac.rs

Датум последње измене: 16.10.2024.

# Велики језички модели

Место и структура великих језичких модела

# Место великих језичких модела

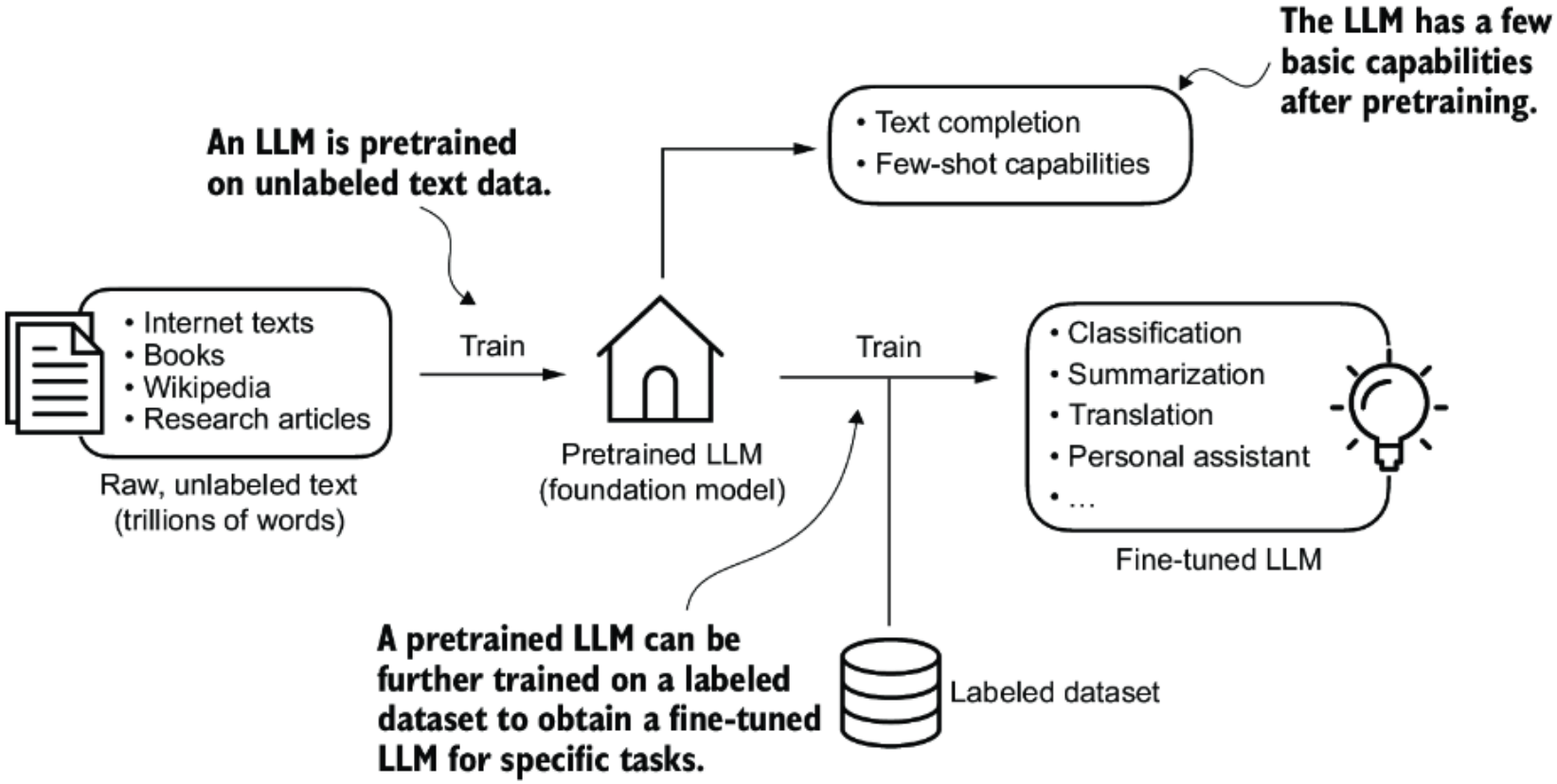


# Велики језички модел – подаци за тренинг

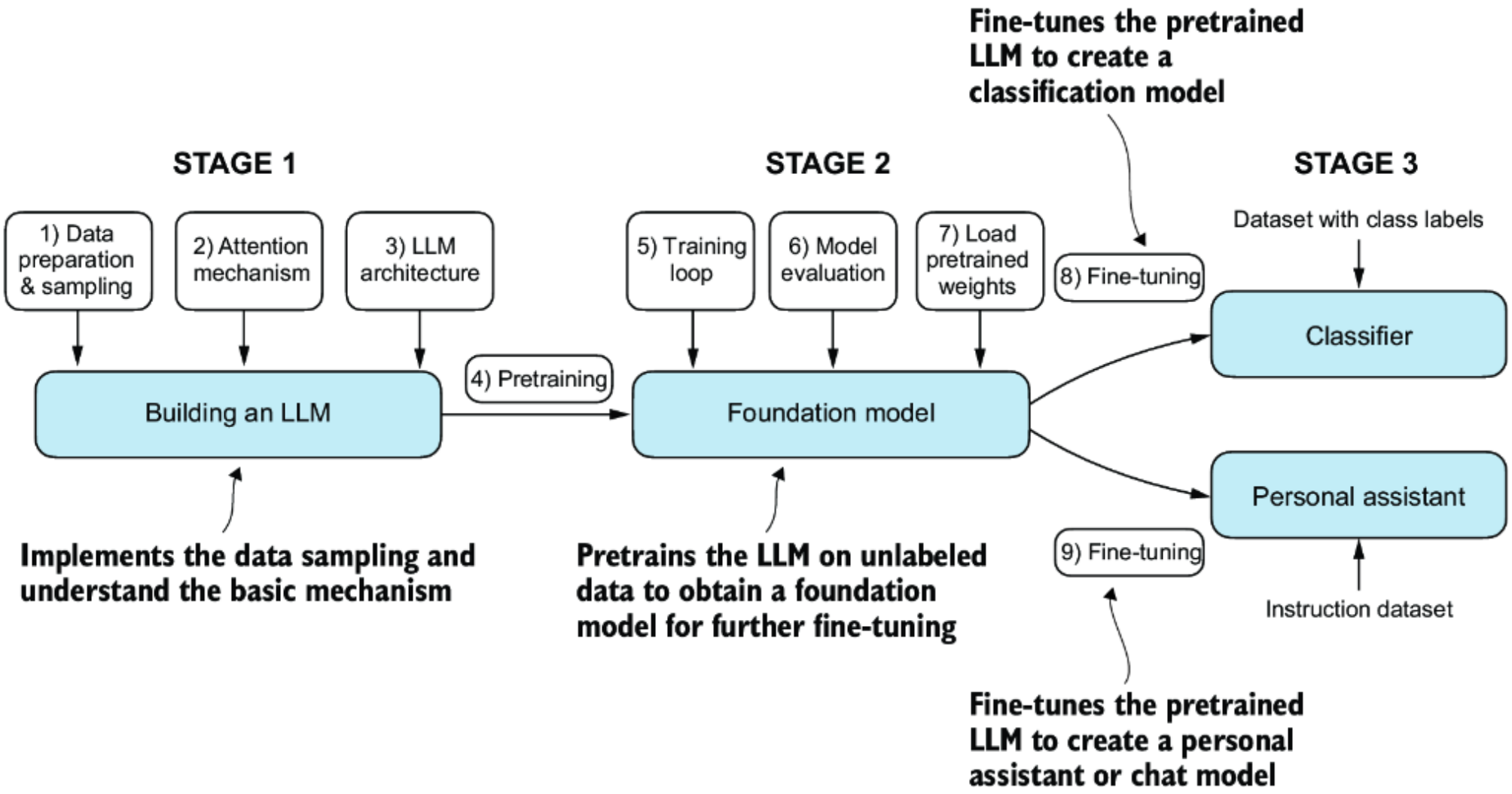
Table 1.1 The pretraining dataset of the popular GPT-3 LLM

| Dataset name | Dataset description | Number of tokens | Proportion in training data |
|---|---|---|---|
| CommonCrawl (filtered) | Web crawl data | 410 billion | 60% |
| WebText2 | Web crawl data | 19 billion | 22% |
| Books1 | Internet-based book corpus | 12 billion | 8% |
| Books2 | Internet-based book corpus | 55 billion | 8% |
| Wikipedia | High-quality text | 3 billion | 3% |

# Фазе креирања и коришћења LLM



# Фазе креирања и коришћења LLM - детаљно



# Фазе у креирању LLM

1) Припрема података
   - Токенизација – дели се улаз у токене на смислен начин.
   - Узорковање података – узима се узорак улазних података и припремају се за фазу обуке, обично раздвајањем скупа података на реченице одређене дужине и генерисањем очекиваног одговора.
   - Уметање токена – додељује се сваком од претходних токена у речнику вектор жељених димензија за обуку модела. Свака реч у речнику биће тачка у простору, где је иницијално позиција сваке речи у простору „насумично“ одређена и те позиције буду побољшањ током обуке
   - Током уметања токена ствара се још један слој који представља (у овом случају) апсолутну позицију речи у реченици за обуку. Према томе, реч на различитим позицијама у реченици ће имати различиту репрезентацију (значење).

# Фазе у креирању LLM

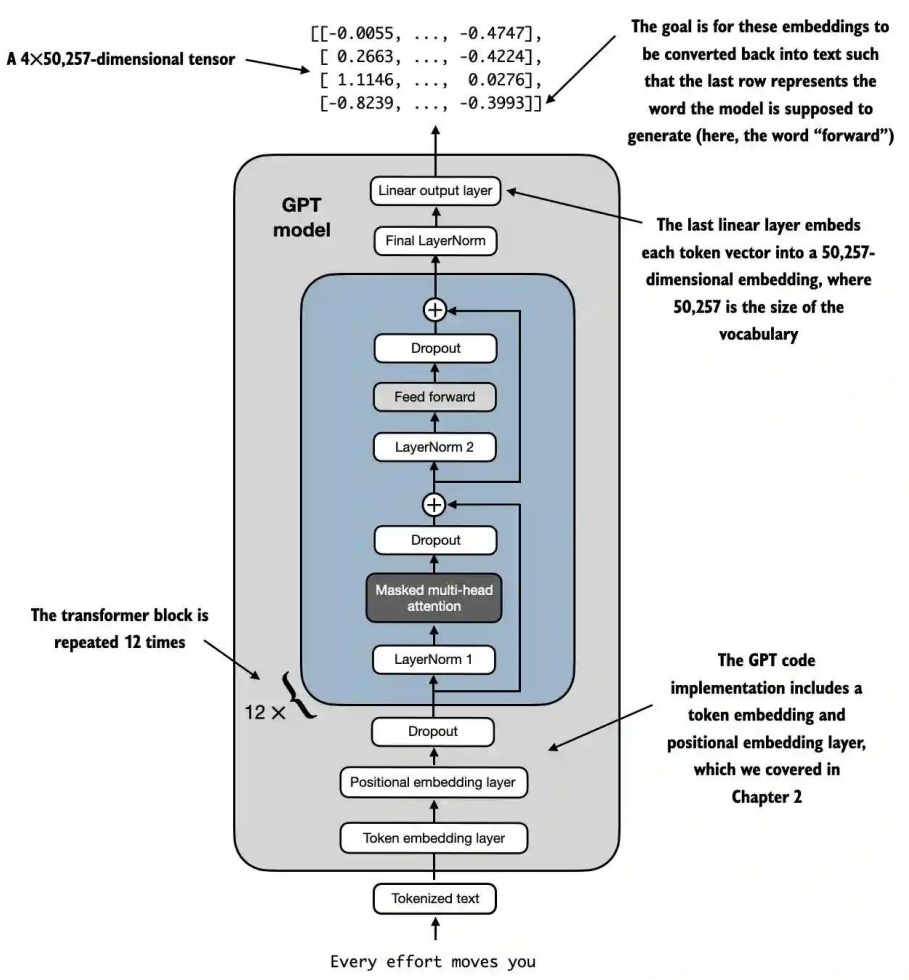
## 2) Механизми пажње

- Механизми пажње омогућавају неуронским мрежама да фокусирају на специфичне делове улаза приликом генерисања сваког дела излаза.
- Они додељују различите тежине различитим улазима и тако користе моделу да одлучи који су улази најрелевантнији за задатак – што је кључно у задацима попут машинског превођења, где је разумевање контекста целе реченице неопходно за тачан превод.
- У традиционалним моделима секвенца-секвенца који се користе за превођење језика, модел кодира улазну секвенцу у векторски контекст фиксне величине. Традицинални приступ се суочава са проблемима са дугим реченицама јер фиксни векторски контекст можда неће ухватити све потребне информације. Механизми пажње решавају ово ограничење омогућавајући моделу да разматра све улазне токене приликом генерисања сваког излазног токена.

# Фазе у креирању LLM

## 3) Архитектура LLM

- Спаја се све заједно, примењују се сви слојеви и креираају се све функције за генерисање текста или трансформацију текста у токене и обрнуто.
- Ова архитектура ће се користити и за обуку и за предвиђање текста након што се заврши обука.



# Фазе у креирању LLM

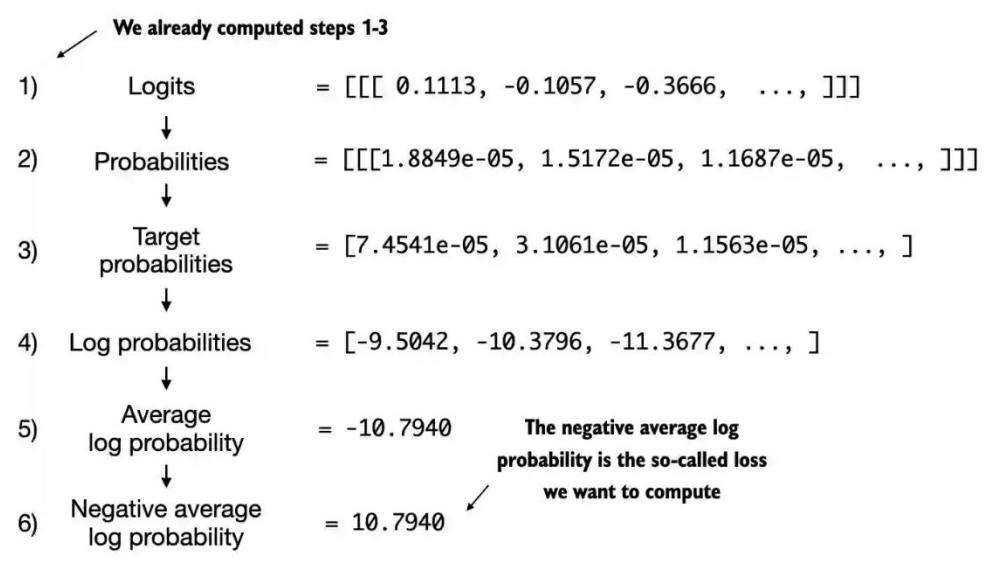
4)-7) Претходно тренирање и тренирање

- Обучити модел од нуле.
- Претходно тренирање је почетна фаза обуке у којој се учи из великог, разноликог скупа података од често милијарди токена.
- Циљ да се развије широко разумевање језика, контекста и различитих врста знања.
- Претходно тренирање је обично изузетно рачунарски скупо и захтева ОГРОМНЕ количине података.
- Често се говори о милионима долара када се обучавају ови модели
- Користи се претходно развијена LLM архитектура где се у циклусима пролази кроз скупове ОГРОМИХ података користећи дефинисане функције губитка и оптимизатор за обуку свих параметара модела.
- Циљ тренирања је да се максимизира вероватноћа исправног токена, што укључује повећање његове вероватноће у односу на друге токене.
- Стога, тежине модела морају бити модификоване тако да је вероватноћа максимизирана.

# Фазе у креирању LLM

## 4)-7) Претходно тренирање и тренирање

- Ажурирање тежина се врши путем пропагације уназад. Ово захтева функцију губитка да би се максимизирала. У овом случају, функција ће бити разлика између извршеног предвиђања и жељеног.



# Фазе у креирању LLM

## 8)-9) Фино подешавање

- Ту се узима већ претходно обучени модел и даље се обучава на специфичнијем скупу података.
- Овај скуп података је обично мањи и фокусиран на одређени домен или задатак.
- Сврха финог подешавања је прилагођавање модела како би се боље понашао у специфичним сценаријима или на задацима који нису били добро покривени током претходног-тренинга.
- Ново знање додато током финог подешавања се више односи на побољшање перформанси модела у специфичним контекстима, а не на ширење његовог општег знања.

# Фазе у креирању LLM

## 8)-9) Фино подешавање

- Фино подешавање великих модела мења само мали део модела, смањује број параметара које треба обучити, чиме се штеди меморија и рачунарски ресурси.
- То је зато што:
  - Смањује се број параметара који се могу обучавати, што чини обуку бржом и захтева мање меморије.
  - Уместо да се потпуно ажурира тежине слоја (матрице), матрица се апроксимира као производ 2 мање матрице, смањујући ажурирање.
  - Одржава оригиналне тежине модела непромењеним и само се ажурирају мале «нове матрице».
  - Оригинално знање модела се чува, а прилагођава се само  оно што је неопходно.
  - Након финог подешавања, уместо да се чува «нови модел» за сваки задатак, потребно је да се чувају само «нове матрице», које су веома мале у поређењу са целим моделом.