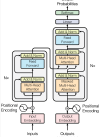


Self attention and Transformers

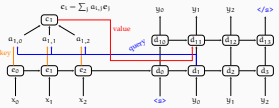
- RNNs have been the primary architecture for modeling sequences. However,
 - RNNs condition only on preceding input
 - RNNs are inherently sequential – difficult to parallelize
- Transformers are designed to resolve these problems
- They use rely mainly on a form of attention, called *self attention*

Transformer architecture



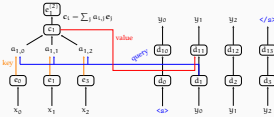
- An encoder-decoder architecture only using self attention
- The encoder is a stack of blocs consisting of self attention followed by a feed-forward layer
- The decoder is similar but,
 - Future input is 'masked'
 - Besides self attention, it also attends to the encoder states
- Both encoder and decoder are stacked
- Both of them use 'multi-head' attention

Sequence-to-sequence RNN with attention



First step towards Transformers

What if we drop the recursion?



Calculating the context vector

- The context vector is the sum of the encoder states, weighted by attention, $a_{1,j}$

$$c_1 = \sum_j a_{1,j} e_j$$
- Typically the weights are normalized through *softmax* the result is an attention distribution

$$a_{1,j} = \frac{e^{f(d_1, e_j)}}{\sum_k e^{f(d_1, e_k)}}$$
- The attention function, $f(\cdot)$ computes the relevance of encoder state e_j to the decoder state d_1

The attention function

- The attention function returns a high value if the encoder state (the key) is relevant to the decoder state (the query)
- Most commonly, the attention function is a version of dot product between the query and the key

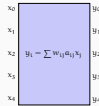
$$f(d_i, e_j) = d_i^T e_j$$

$$f(d_i, e_j) = d_i^T W_d e_j$$

$$f(d_i, e_j) = \frac{d_i \cdot e_j}{\sqrt{d}}$$

Self attention

a simplified first view



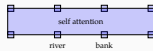
- The input to the next layer is a weighted sum of all previous inputs
- The attention ($a_{1,j}$) depend on the inputs x_i
- The general idea is to assign higher weights to inputs that help predict the output
- Similar to convolution, but we are looking at the whole input range
- No recurrence
- The output is permutation invariant (order of the inputs is not important)

Simplified self attention

$$a_{ij} = \frac{e^{x_i \cdot x_j}}{\sum_k e^{x_i \cdot x_k}} \quad y_i = \sum_j a_{ij} x_j$$

- Dot product weights the inputs with high similarity
- We do not necessarily want similarity

An example



- We want representation of 'bank' to attend to 'river'
- But their embeddings are not likely to be similar
- Dot product by itself is not useful
- Note also the similarity with convolutions, but with unbounded distance

Self attention

typical approach

- Each input is transformed to three different vectors
 - $k_i = W_k x_i$ (key)
 - $q_i = W_q x_i$ (query)
 - $v_i = W_v x_i$ (value)
- The transformations (W_k, W_q, W_v) are learned
- The attention function becomes

$$\text{softmax}\left(\frac{q \cdot k}{\sqrt{d}}\right) v$$
- This can be computed efficiently, and in parallel

Multi-head attention

- In practical implementation multiple attention functions are learned in parallel
- conceptually, this learns different type of relations between input units

How about information from the sequence?

- Self attention is not sensitive to the order of input elements
- The solution is encode the position information on each embedding
- Two common approaches
 - Position encodings: combination of harmonic (cosine, sine) functions with different periods
 - Recent systems also include relative positional embeddings

Positional Encoding

Transformer Encoder

C. Gillekens, NBI / University of Tilburg

Transformer Seminar 2024

12 / 19

Transformer architecture

Output Projection

- The first layer is an *embedding* layer: no information from context
- Subsequent layers use attention followed by a non-linear transformation (feed-forward layer)
- Feed-forward layer is a projection an up-projection followed by projection back to input/output dimensions
- Input and output dimensions to each Transformer block is the same
- Layer normalization is after (sometimes before) the attention and feed-forward calculations

C. Gillekens, NBI / University of Tilburg

Transformer Seminar 2024

13 / 19

Transformer architecture: attention

- All queries, keys and values are combined to matrices
- Instead of individual dot product, matrix multiplications is used

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

- On the decoder side, future input is masked
- The result of matrix multiplications are normalized layerwise

C. Gillekens, NBI / University of Tilburg

Transformer Seminar 2024

14 / 19

Transformer architecture: multihead

- All attention layers are multiplied
- Conceptually, each head learns a different property of the input to 'attend to'

C. Gillekens, NBI / University of Tilburg

Transformer Seminar 2024

15 / 19

Feed-forward layer

- The output of the attention module is passed to a feed-forward network (FFN)
- Typically FFN first applies a non-linear projection to a higher dimension (typically 4 times its input), and projects it back to the input/output (model) dimension
- Conceptually,
 - attention decides what lower-layer tokens in the sequence to use as input
 - the FFN transforms (non-linearly) to a new representation

Feed-Forward Network

C. Gillekens, NBI / University of Tilburg

Transformer Seminar 2024

16 / 19

Layer normalization

- Normalization helps keeping activations (and gradients) in a reasonable range
- This is important for practical reasons: most optimizers work better if activations do not have extreme values
- Original transformers uses post-norm: normalization is done after attention/FFN calculations
- Most recent models use pre-norm
- 'Layer Norm' in original Transformers simply calculates the z-score over the activations (features)
- A common alternative is RMSNorm
- Another common option is batch-norm: normalize across the batch rather than feature dimension

Layer Normalization

C. Gillekens, NBI / University of Tilburg

Transformer Seminar 2024

17 / 19

Attention in decoder

- The decoder has to decode tokens without the knowledge of the true tokens
- The attention on the decoder is 'masked', each token can only attend to the preceding tokens

$$\text{mask}\left(\text{softmax}\left(\frac{q \cdot k}{\sqrt{d}}\right)\right)v$$

Masked Attention

C. Gillekens, NBI / University of Tilburg

Transformer Seminar 2024

18 / 19

Some closing notes

- The Transformer architecture is the main architecture used in modern (large) language models (with minor changes)
- Original Transformer is an encoder-decoder model, used for machine translation
- Modern language models can be either
 - Encoder only
 - Encoder-decoder
 - Decoder only
- Reading: Jurafsky and Martin (2026, Chapter 9)

Next:

- Transformer language models
- Reading: Jurafsky and Martin (2026, Chapters 10 & 11)

C. Gillekens, NBI / University of Tilburg

Transformer Seminar 2024

19 / 19

Additional reading, references, credits

- The diagrams of the Transformer architecture is from the original Transformers paper (Vaswani et al., 2017)

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

277

278

279

280

281

282

283

284

285

286

287

288

289

290

291

292

293

294

295

296

297

298

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

320

321

322

323

324

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

384

385

386

387

388

389

390

391

392

393

394

395

396

397

398

399

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

496

497

498

499

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543

544

545

546

547

548

549

550

551

552

553

554

555

556

557

558

559

560

561

562

563

564

565

566

567

568

569

570

571

572

573

574

575

576

577

578

579

580

581

582

583

584

585

586

587

588

589

590

591

592

593

594

595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648

649

650

651

652

653

654

655

656

657

658

659

660

661

662

663

664

665

666

667

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702

703

704

705

706

707

708

709

710

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756

757

758

759

760

761

762

763

764

765

766

767

768

769

770

771

772

773

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

810

811

812

813

814

815

816

817

818

819

820

821

822

823

824

825

826

827

828

829

830

831

832

833

834

835

836

837

838

839

840

841

842

843

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863

864

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

918

919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

C. Gillekens, NBI / University of Tilburg

Transformer Seminar 2024

A.1