Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model

05 July 2022

Narutatsu (Edward) Ri

Motivation

Goal: Understand the detailed process behind language comprehension in the brain

- Neuroimaging research isolates particular linguistic computations in controlled setting and mapped them onto brain activity (Bookheimer, 2002; etc.)
    - Limited generalizability
    - Difficult to create a holistic model that reflects full complexity of natural language
- Transformers (Vaswani, 2017) show success of capturing sophisticated representations of linguistic structure (Manning et al., 2020; Linzen & Baroni, 2021)

Can we use Transformers to model human brain activity during natural language comprehension?

The Transformer and BERT

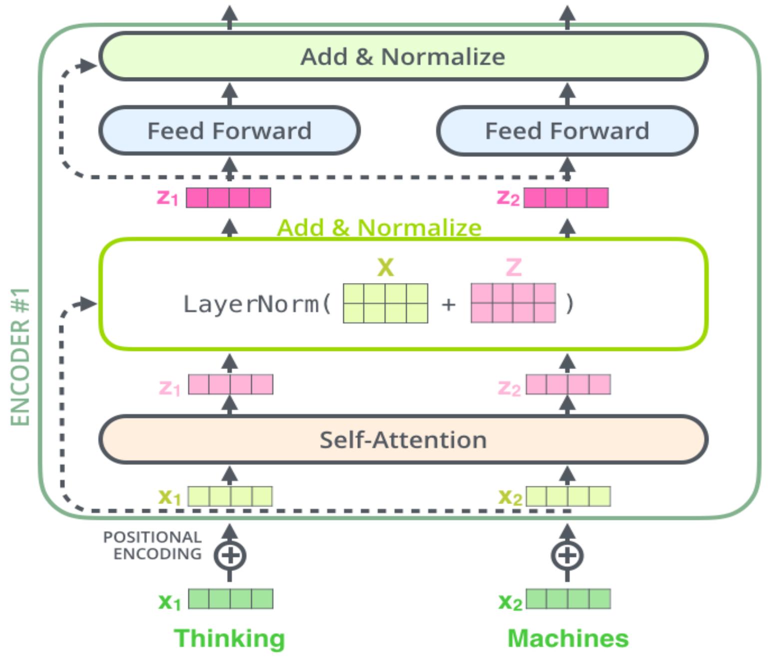


Fig. 1: Encoder Architecture of Transformer
(The Illustrated Transformer)

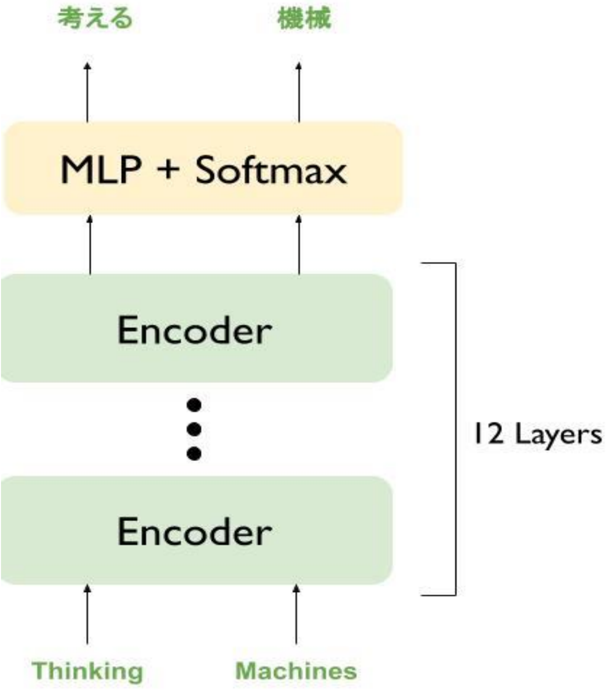


Fig 2: BERT’s Architecture

Transformer Encoder

- Bidirectional (Decoder is unidirectional)
- Convert input words to static embedding (e.g. one-hot encoding) and add positional encoding
- Apply self-attention (12 individual attention heads for standard model)
- Feed to MLP

BERT

- Stack Encoder part of Transformer (12 layers) followed by an output layer (interchangeable with different layers for specific downstream task)
- Index of elements with high values correspond to certain symbols, which is considered the output

Embeddings and Transformations

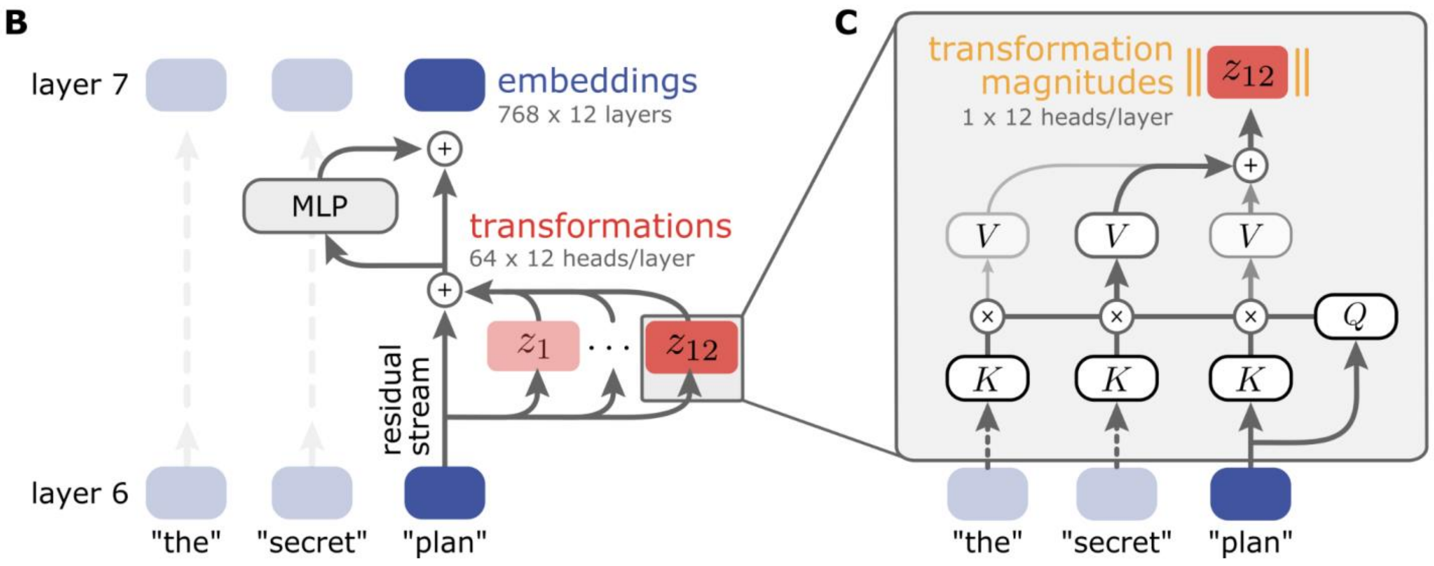


Fig. 3: Embeddings vs. Transformations. Transformations are added to embeddings

Embeddings

- "Residual stream" (Elhage, 2021)
  - Model's internal representation of linguistic content
  - Embeddings accumulate previous computations/information over layers
- Majority of previous work focus on attempting to find relationship between embeddings and neural activity

Transformations

- Localized computations
  - The transformation added to the embedding from the previous encoder layer
  - Can be broken down into independent attention heads (12 for standard model)
  - Individual heads shown to have functional specialization where particular heads approximate particular linguistic operations (Clark, 2019)
- Not much preceding work done

(Not directly relevant to the original paper): Why the transformations are considered local representations, when the transformations are still obtained by applying computations on the incoming embeddings?

Work Summary

Primary argument: internal computations implemented by the functionally-specialized attention heads provide a more direct window onto linguistic processing in the brain than embeddings

1. Used encoding models to evaluate how well different language model classes predict fMRI data acquired during natural language comprehension
2. Examined performance patterns across different layers
    1. Transformations better recapitulate the cortical processing hierarchy across language areas
3. Decomposed transformations into individual attention heads
    1. Correlation observed between performance on predicting brain activity and predicting syntactic dependencies
    2. Find that certain properties of these heads fall along gradients in a low-dimensional cortical space

Predicting Brain Activity from Language Models

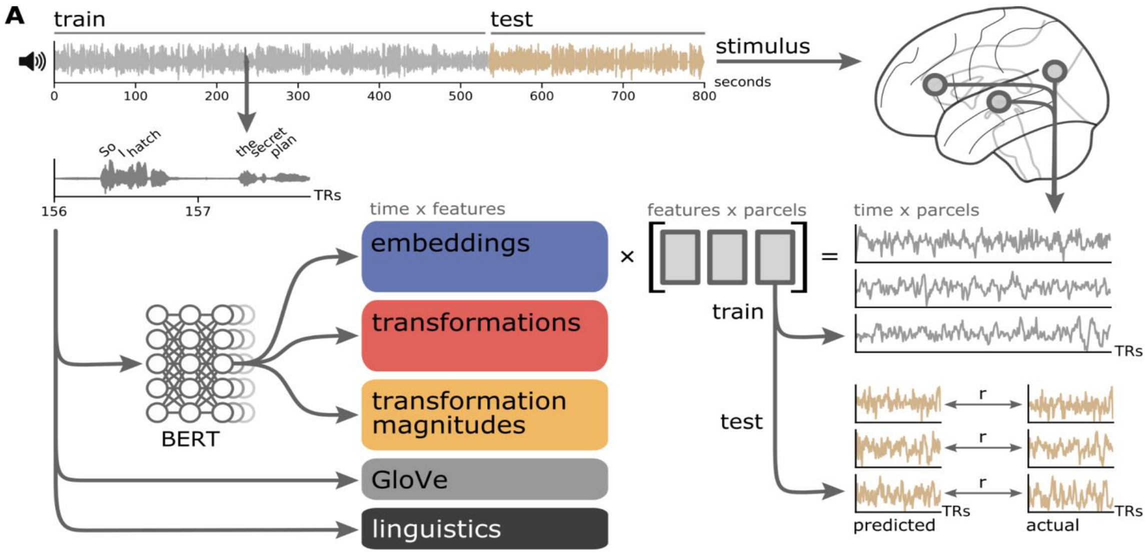


Fig. 4: Encoding models for predicting brain activity from language models

- Predict brain activity from language models with encoding model
  - Dataset is audio of spoken stories (two story datasets)
    - “slumlord”: “Reach for the Stars One Small Step at a Time”
      - 18 subjects (18~27 years), 13 minutes (550 TRs), 2,600 words
    - “black”: “I Knew You Were Black”
      - 45 subjects (18~53 years), 13 minutes (534 TRs), 1,500 words
  - BERT model
    - BERT-base-uncased
    - Standard pretrained model, no task-specific finetuning

Predicting Brain Activity from Language Models

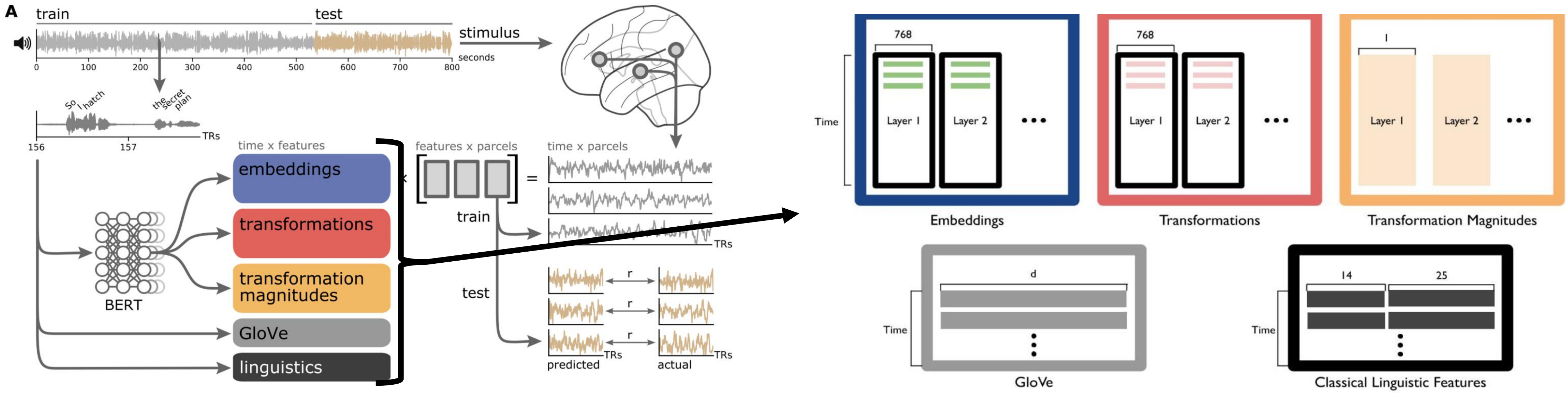


Fig. 5: Structure of representations for each language model class

- Obtain linguistic state (time x features) at each time step for five different language model classes
  - Classical linguistic features (14 + 25)
    - POS (14)
    - Dependency relations (25)
  - Static embeddings: GloVe (Not specified, 50~300)
  - Contextual embeddings: BERT (768 x 12 layers)
  - Transformations (64 x 12 heads x 12 layers)
  - Transformation magnitudes (1 x 12 heads x 12 layers)
- Embeddings and transformations are concatenated across all layers
- Banded ridge regression to learn weights (features x 192 parcels) to map to measured brain activity (time x 192 parcels)

Predicting Brain Activity from Language Models

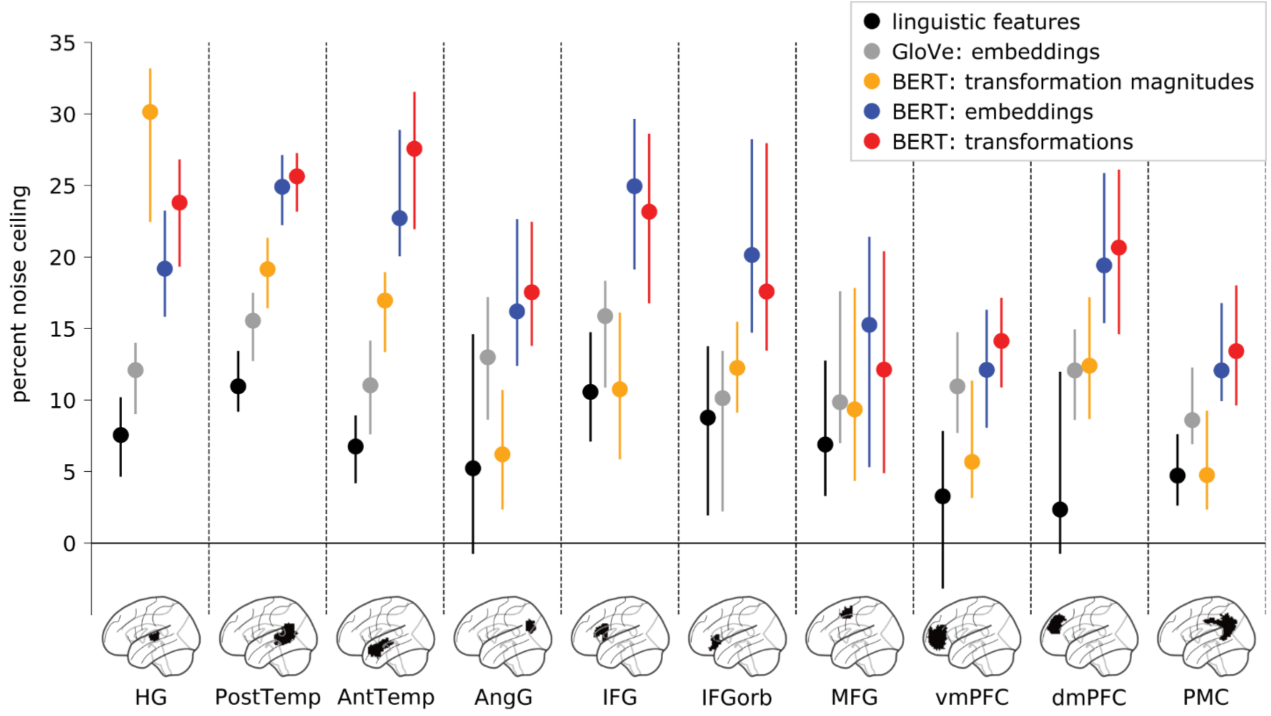


Fig 6: Comparing five classes of language models across cortical language areas

- Embeddings and transformations outperform linguistic features and static embeddings in most language ROIs
- Transformation magnitudes outperform static embeddings and linguistic features in lateral temporal areas but not in higher-level language areas
- Transformations roughly match embeddings across all ROIs

Predicting Brain Activity from Language Models

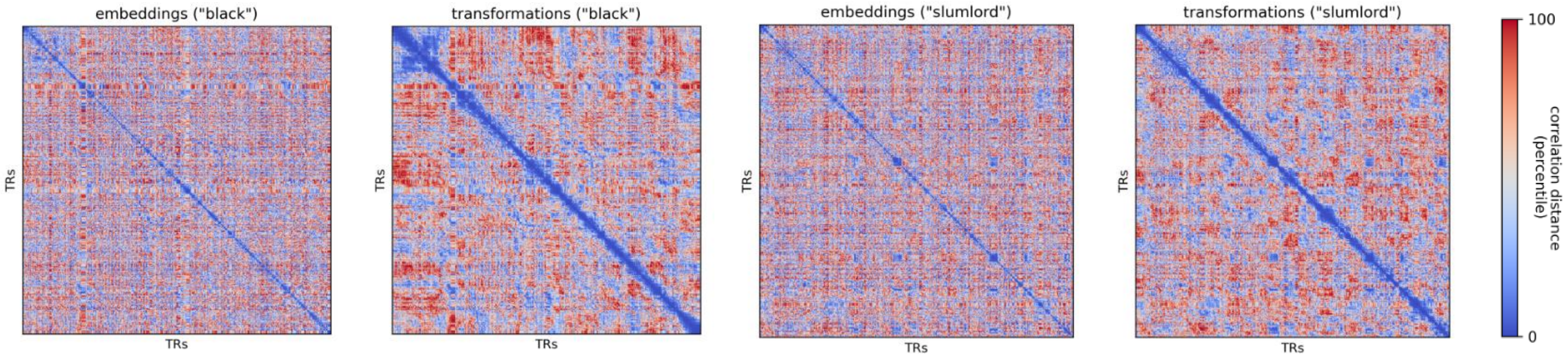


Fig 7: TR-by-TR representational dissimilarity matrices (RDMs) concatenated across all layers

Embeddings and transformation representations are fundamentally different

- Average TR-by-TR correlation between embeddings and transformations for both datasets is effectively zero (-.004 ± .009 SD)
- Embeddings and transformations yield visibly different TR-by-TR representational geometries (Fig. 7)
    - “Visibly different”: Not quite convinced...
    - What is the significance of this?

Predicting Brain Activity from Language Models

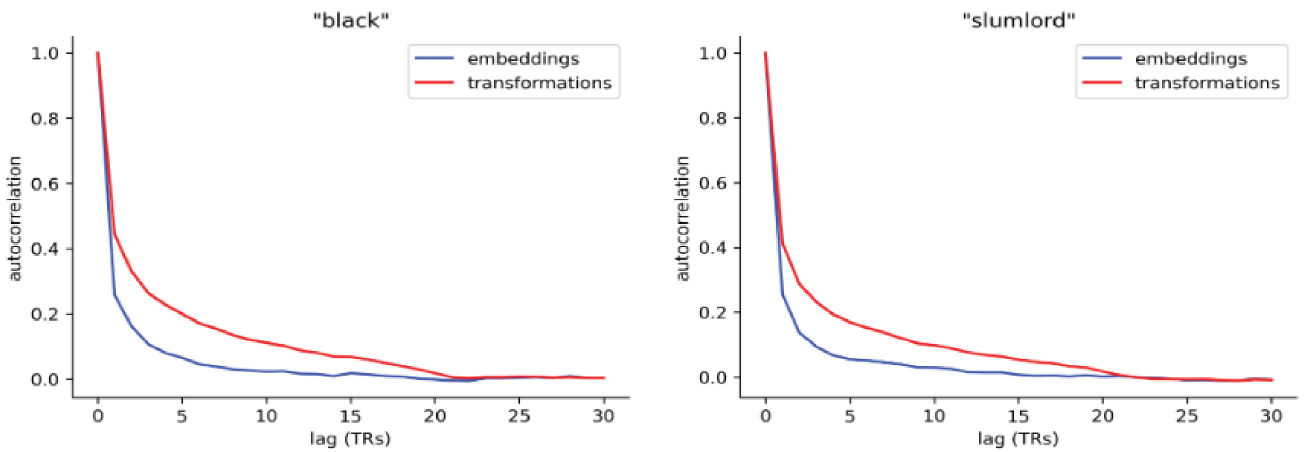


Fig 8: Transformations have higher autocorrelation than embeddings in both stimuli

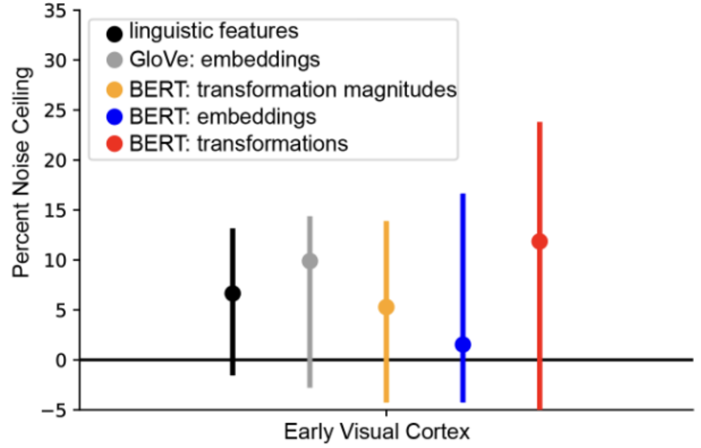


Fig 9: Encoding performance for three classes of language models

- Transformations have considerably higher temporal autocorrelation than the embeddings (Fig. 8)
  - What is the significance of this?
- Control analysis: Evaluated features in a non-language ROI (early visual cortex) and found that no models captured a significant amount of variance
  - We expect BERT's features to perform better for ROIs responsible for language modeling because we assume BERT's features are more robust than previous classes
  - Thus, showing that performance levels for visual cortex is relatively the same for all classes reinforces the argument that certain ROIs are responsible for language modeling?

Embeddings vs. Transformations: Similarity Across Layers

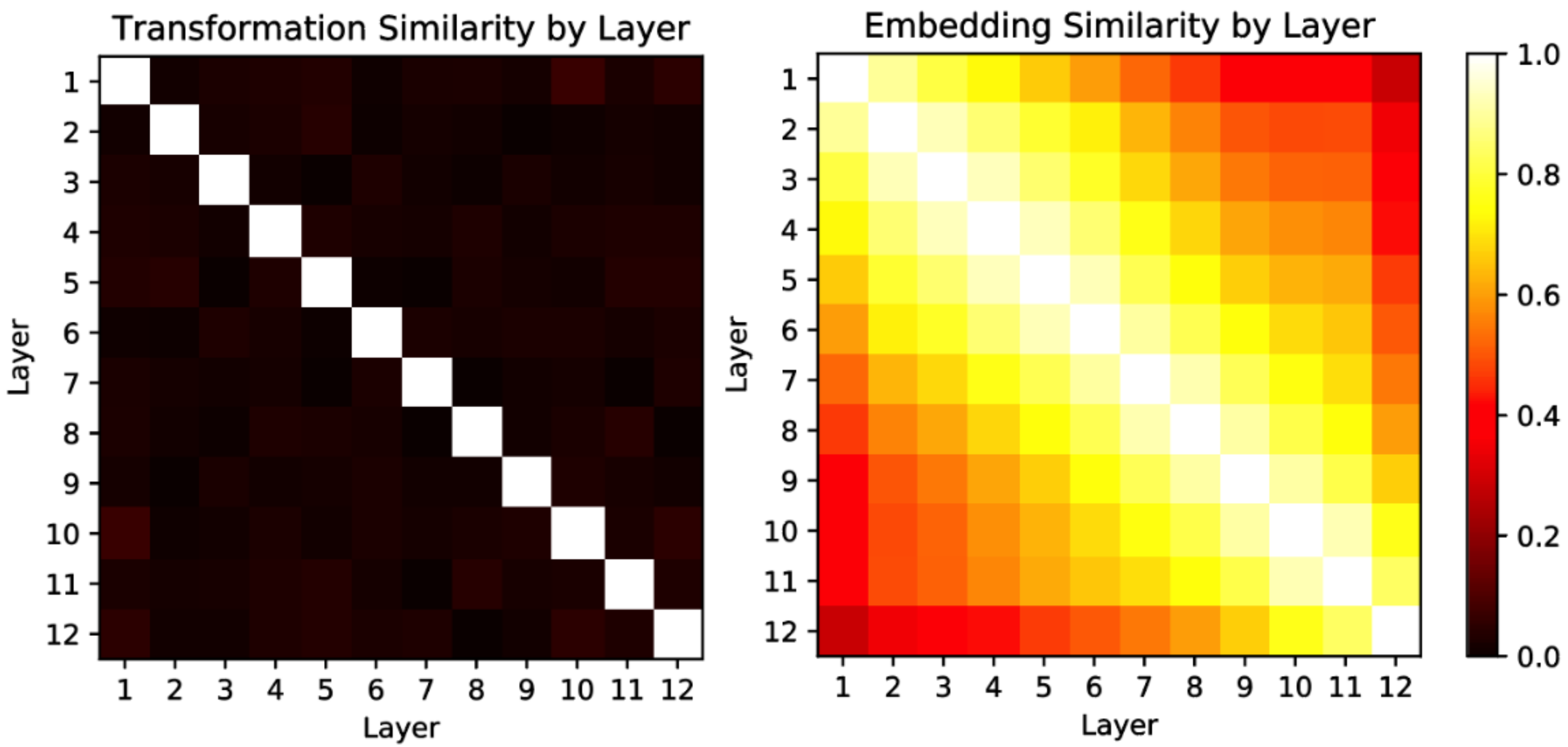


Fig 10: Similarity between transformations and embeddings across layers

- Embeddings are increasingly contextualized; later layers reflect more complex linguistic relationships (Tenney et al., 2019)
- Transformations are largely independent from layer to layer (Fig. 10)

Embeddings vs. Transformations: Similarity Across Layers

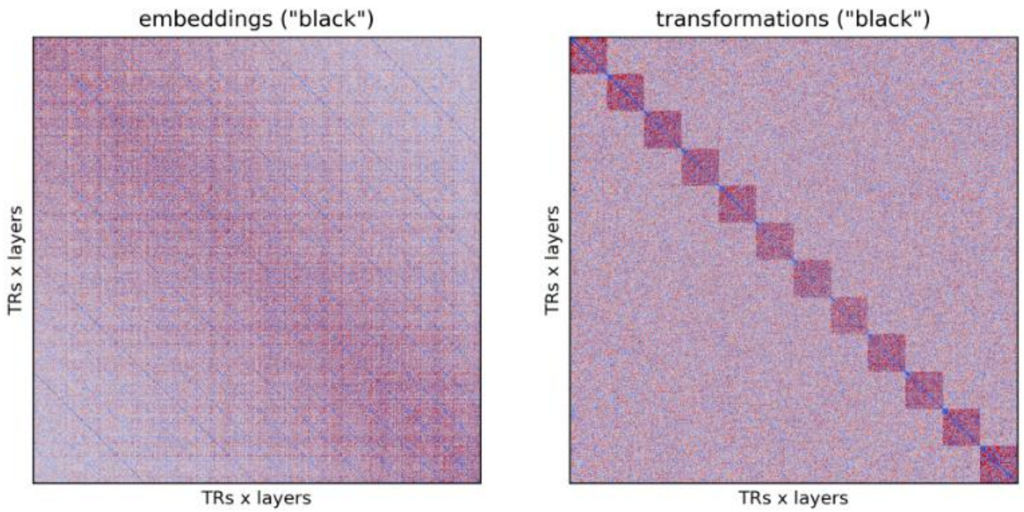


Fig 11: TR-by-TR RDMs within and across each layer of BERT (TR-by-TR RDM x 12 layers)

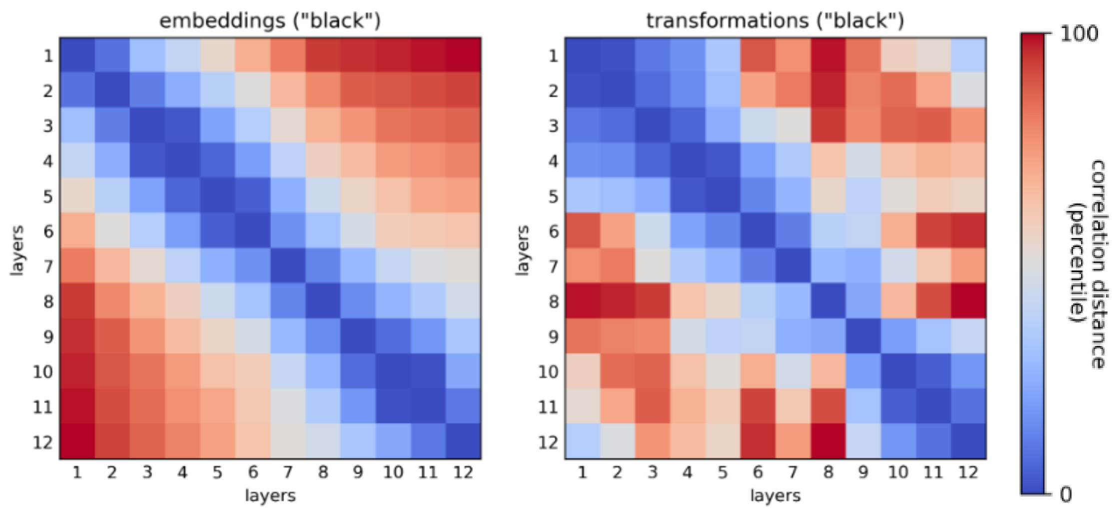


Fig 12: Second-order layer-by-layer representational geometry of TR-by-TR RDMs

- Transformations produce more layer-specific representational geometries
- Fig. 11: Comparison of representation similarities between layers by TR
  - Faint blue diagonal lines visible for embeddings: Similarity between embeddings of different layers at same time point (also shown in previous slide)
- Fig. 12: Comparison of TR-by-TR RDM similarities between layers
  - Demonstrate that layer-wise representational geometries evolve sequentially across layers
  - Layer 1 and 12 show similarity between RDMs; Layer 1 and 8 show largest difference
    - Could be some pattern, could be by "chance?"

Layer-wise Encoding Performance: Comparison

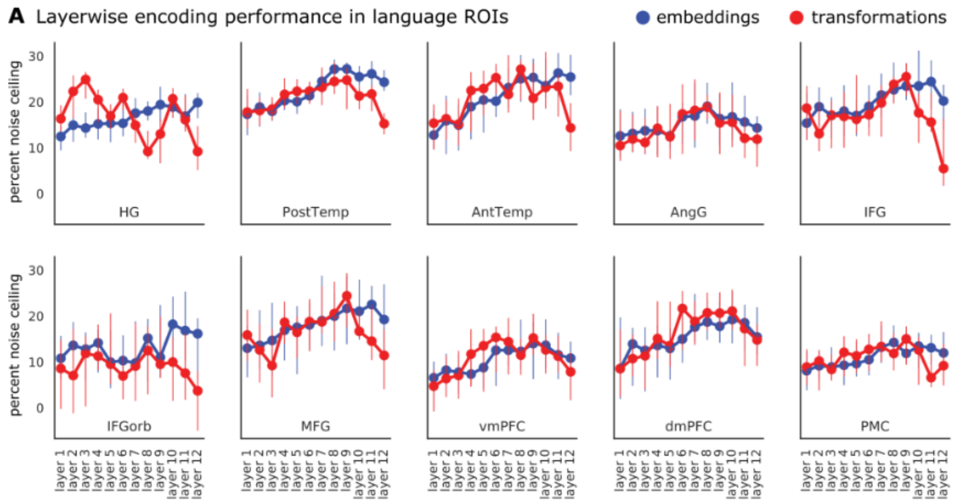


Fig 13: Layer-wise model performance in ten left-hemisphere language ROIs

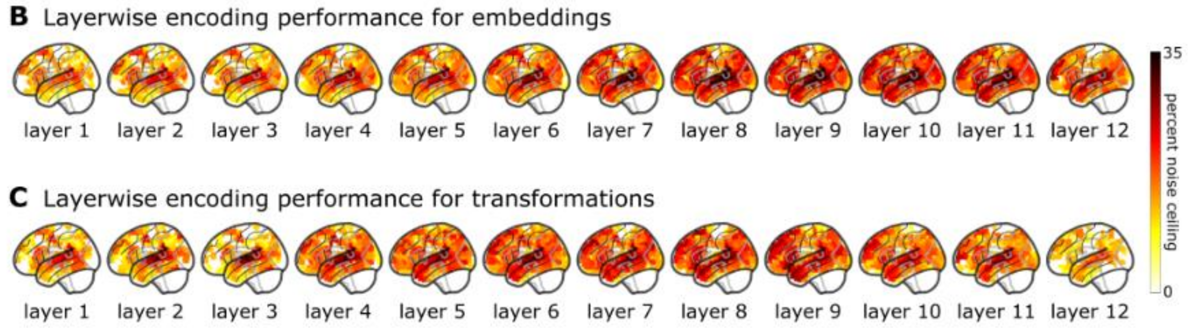


Fig 14: Model performance for each layer across all cortical parcels

- Performance of embeddings increased roughly monotonically across layers, peaking in late-intermediate or final layers (Fig. 13)
  - Observed across most ROIs, suggesting that the hierarchy of layer-wise embeddings does not cleanly map onto the cortical hierarchy for language comprehension
    - Why so? Because we would expect to see similar performances across ROIs if a clean mapping exists?
- Performance of transformations yield more layer-specific fluctuations
  - Suggesting that computations implemented at particular layers map onto brain in a more specific way than embeddings
- Beyond language areas, similar pattern is observed for cortical parcels
  - What is the significance of this? I thought we established earlier language models do not capture functions of other non-language ROIs well?

Layer-wise Encoding Performance: Comparison

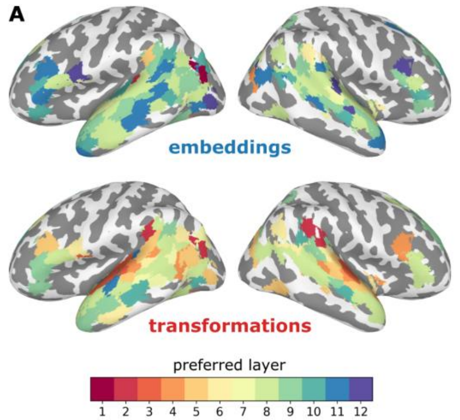


Fig 15: Layer preferences are visualized on the cortical surface

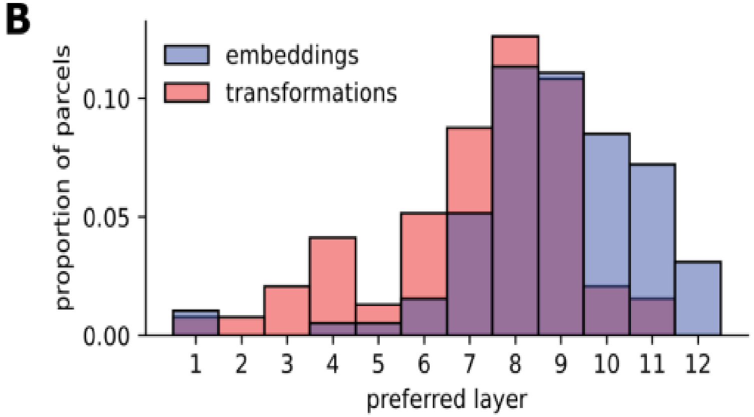


Fig 16: Histogram of preferred layer across cortical parcels in language ROIs

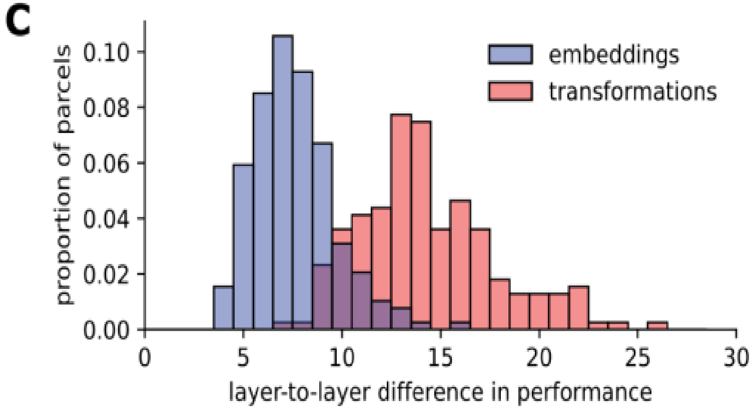
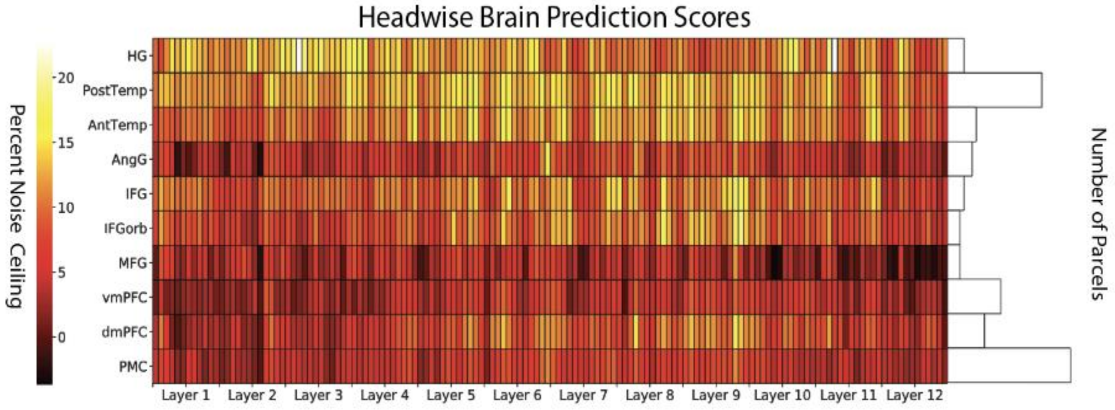


Fig 17: Distribution of magnitude of layer-to-layer differences in encoding performance

- Visualized which layer yielded the peak performance for a given cortical parcel (Fig. 15)
- Average performance for embeddings peaked significantly later than performance for transformations (Fig. 16)
  - Across language parcels, performance for transformations peaks at intermediate layers, while performance for embeddings peaks in later layers
    - What is the difference in argument between Fig. 16 and Fig. 13?
- Quantified magnitude of difference in predictive performance from layer to layer for all cortical parcels (Fig. 17)
  - Found that transformations have larger differences in performance between neighboring layers
  - How to interpret this figure? What is horizontal axis?
  - Computations implemented by transformations are considerably more layer-specific than embeddings

Interpreting Transformations via Head-wise Analysis



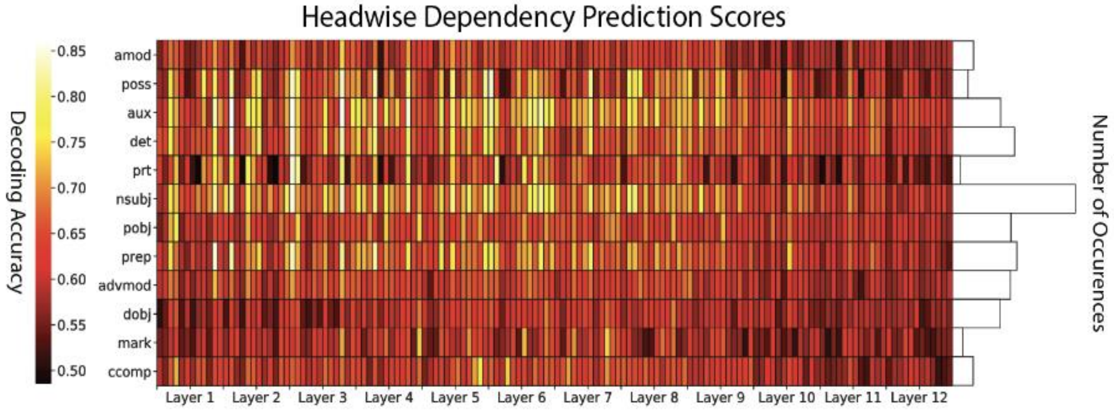


Fig 18: Head-wise brain prediction scores and dependency prediction scores

- Classical linguistic features are poor predictors of brain activity and did not generally map onto localized brain regions in the context of naturalistic narratives (Fig. 6)
- Identified brain prediction scores for head-wise transformations (Fig. 18 left)
- Identified which attention head predicts classical syntactic dependency (Fig. 18 right)
  - Example: Layer 6, Head 11 best predicts direct object

Interpreting Transformations via Head-wise Analysis

- Head most associated with a given dependency generally outperformed the dependency itself (Fig. 19)
- Dense, emergent head-wise transformations are better predictors of brain activity than sparse, classical linguistic indicator variables
    - Head-wise transformations are considerably higher-dimensional (64 dimensions) than the corresponding one-dimensional dependency indicators
        - Head-wise transformations have richer expressive capabilities
        - Not exactly sure how these indicators are represented
- After reducing head-wise transformations that best predicts corresponding dependency to a single dimension, one-dimensional transformation still better predicts brain activity than the dependency itself (Fig. 20)
    - Transformations do not simply indicate presence of syntactic dependency, but rather capture an approximation of the direct object relationship in the context of the ongoing narrative

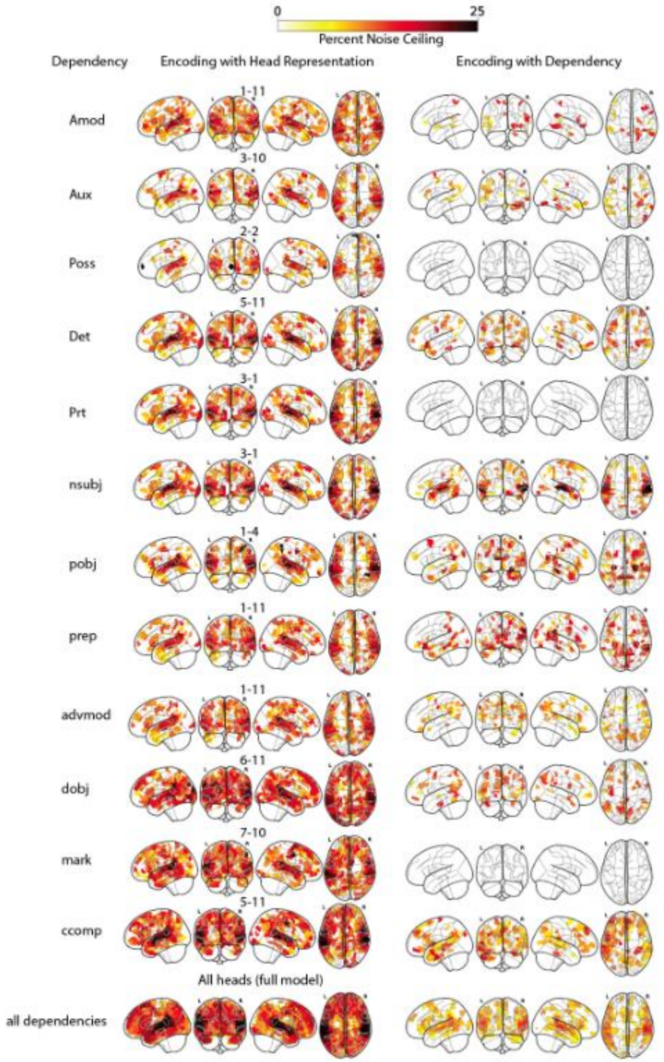


Fig 19: Comparison between encoding performance using head-wise transformation that best predicts syntactic dependency and classical linguistic dependency itself

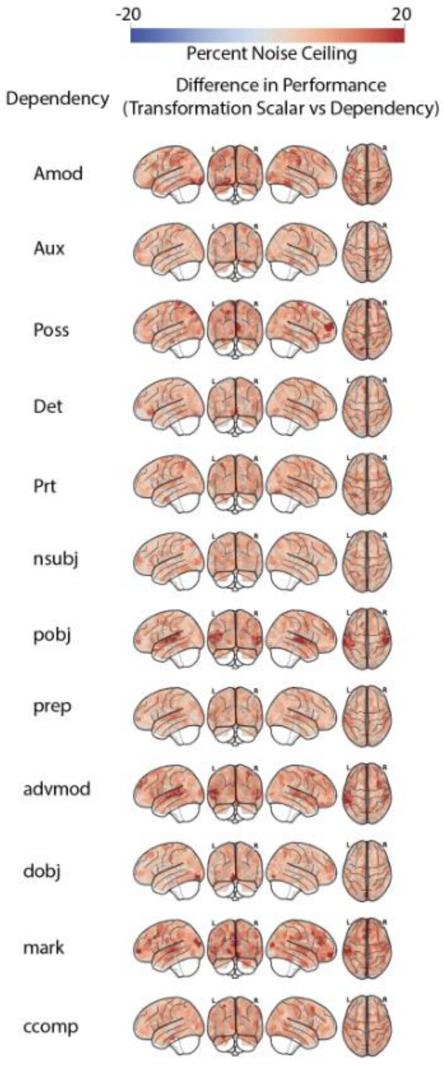


Fig 20: Difference in encoding performance between reduced head-wise transformations and linguistic dependency

Interpreting Transformations via Head-wise Analysis

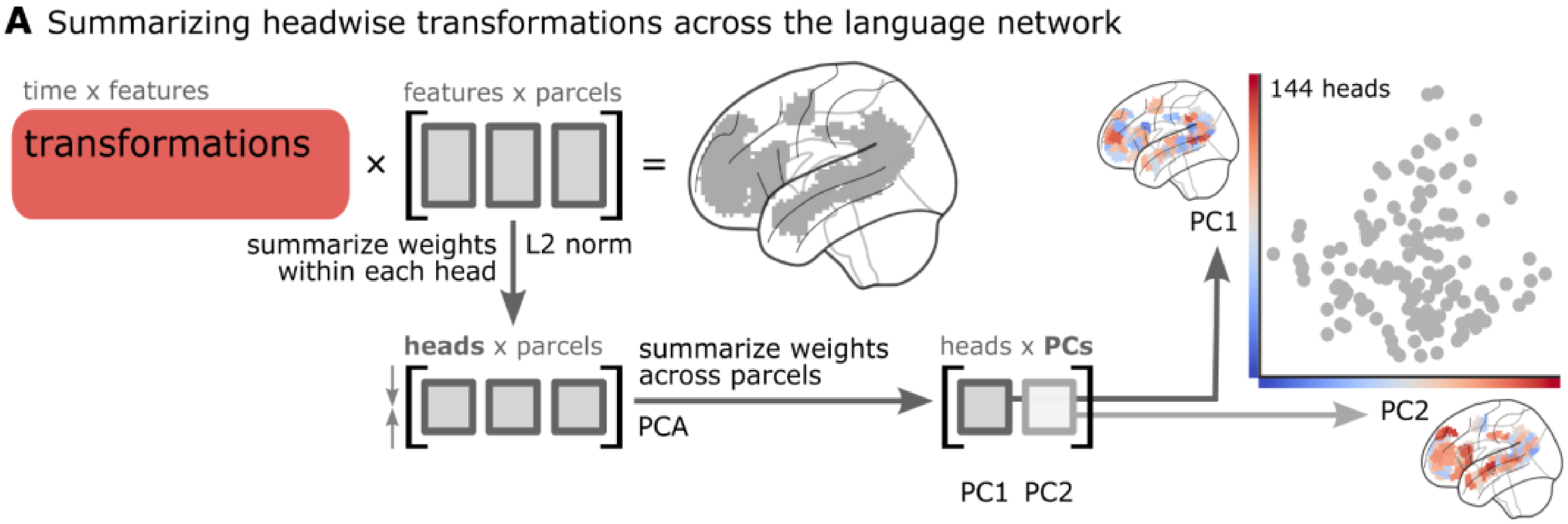


Fig 21: Illustrated process of mapping to lower-dimensional cortical space

- Summarized contributions of all head-wise transformations across the entire language network
  - Segment weight matrix for each parcel into individual attention heads at each layer and computed L2 norm of head-wise encoding weights (Fig. 21)
    - Weight matrix is shaped 9,216 features (64 features × 12 heads × 12 layers) × 192 language parcels
    - Take L2 norm reduces this matrix to 144 heads (12 heads × 12 layers) × 192 language parcels
    - Summarize head-wise weights using PCA, project weights onto first two PCs (90% variance)

Interpreting Transformations via Head-wise Analysis

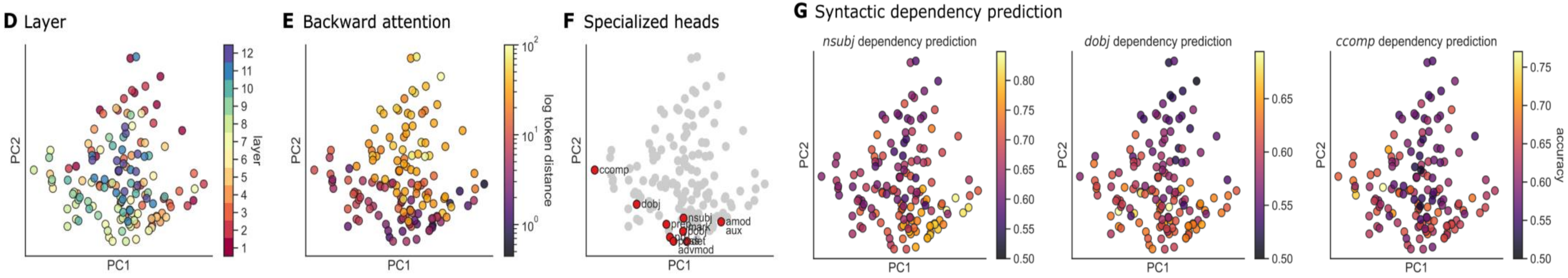


Fig 22: Head-wise transformations in low-dimensional brain space

- Examined structure of "geometry" of head-wise transformations in reduced space
  - Visualized layer numbers of each head and found layer gradient across heads (Fig. 22D)
    - PCs 9, 5, 1 correlated with layer numbers the most (r = 0.45, 0.40, 0.26)
    - Intermediate layers generally in negative quadrant, early and late layers located in positive quadrant
  - Computed average backward attention distance (Fig. 22E)
    - Observed strong gradient of look-back distance increasing along PC2
    - Prefrontal and left anterior temporal parcels correspond to heads with longer look-back distances
  - Functionally specialized heads previously reported in literature (Clark et al., 2019) span PC1 and cluster at negative end of PC2 (Fig. 22F)
    - Corresponding to intermediate layers and relatively recent look-back distance
  - Visualized head-wise dependency prediction scores (Fig. 22G)
    - Observed gradients in different directions
    - Seems like previous literature and current result don't agree with function assignments?

Interpreting Transformations via Head-wise Analysis

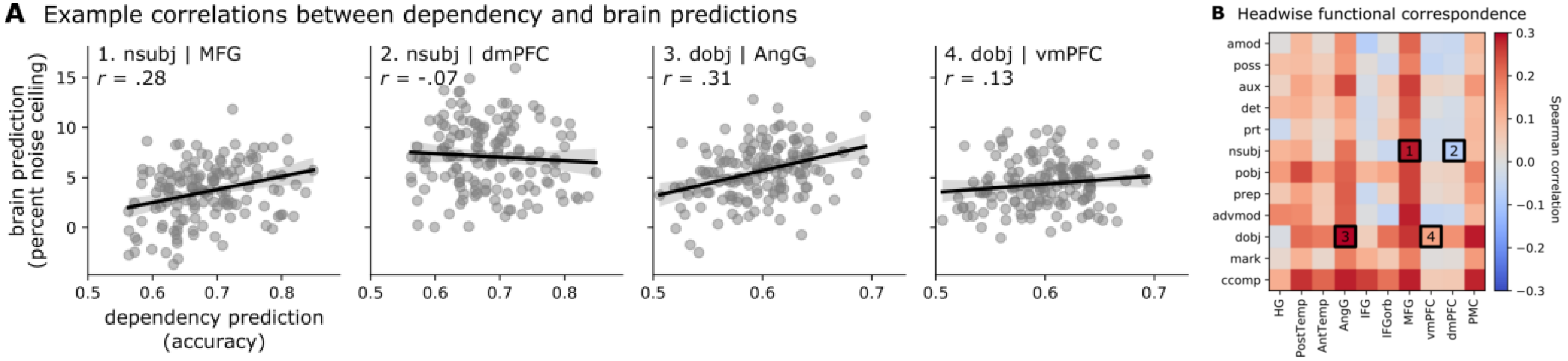


Fig 23: Correspondence between head-wise transformations' brain and dependency predictions

- Quantified correspondence between heads' syntactic information and brain activity prediction performance by computing correlation between brain activity prediction and dependency prediction scores (Fig. 23)
  - i.e. Computed correlation between diagrams in Fig. 18
  - Head-wise correspondence indicate that attention heads containing information about a given dependency also tend to contain information about brain activity for a given ROI, suggesting ROI is involved in computing that dependency
  - Correspondence was high in angular gyrus and MFG across dependencies (Fig. 23B)
    - Observation for MFG is consistent with prior work implicating MFG in both language comprehension and more general cognitive demand (e.g. working memory) (Fedorenko et al., 2011; Mineroff et al., 2018)

Interpreting Transformations via Head-wise Analysis

- From these results, transformations' brain activity prediction performance doesn't correlate too well with classic syntactic dependencies prediction performance
    - Suggests shared information between transformations and certain ROIs may be semantic in nature or reflect contextual relationships beyond the scope of classical syntax
    - Or perhaps something entirely different? Correlation values seem too low for syntactic information to play a significant role in predicting brain activity?
- Despite the formal distinction between syntax and semantics in linguistics, neural computations supporting human language may not cleanly dissociate syntactic and semantic processing
    - Transformer models implicitly learn syntactic operations to produce good linguistic outputs, such structures are generally entangled with semantic content
    - Transformations capture syntactic operations entangled with semantic content, but perhaps transformation magnitudes can help disentangle syntax and semantics
        - Transformation magnitudes reduce transformations down to "activation" of individual heads and might isolate semantic information
            - How so?
        - Insights from NLP (Clark et al., 2019) suggests transformation magnitudes still contain emergent form of syntactic information
        - Transformation magnitudes outperform static embeddings in temporal areas while underperform in angular gyrus, a putative high-level convergence zone for semantic representation

Interpreting Transformations via Head-wise Analysis

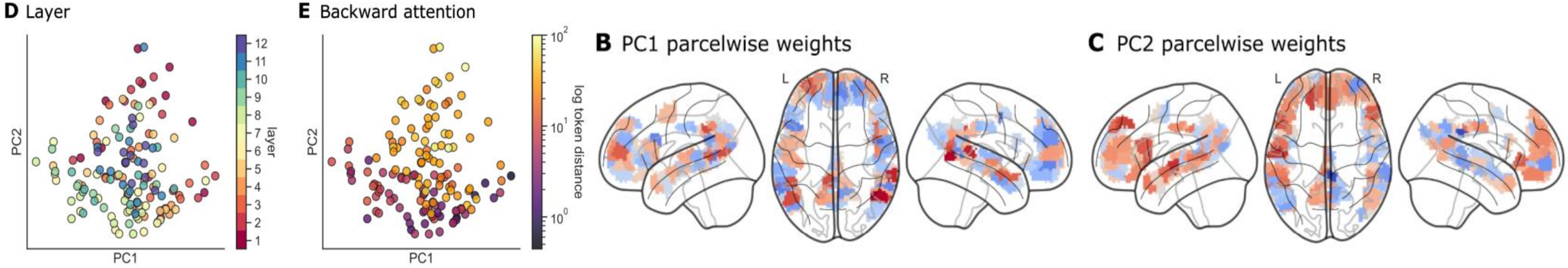


Fig 24: PC1 and PC2 projected back onto the language parcels

- Project PC1 and PC2 back to parcels to obtain weight magnitudes for respective PCs (Fig. 24)
- Functional properties of head-wise transformations map onto certain cortical localization trends
  - Posterior temporal areas assign higher weights to heads at earlier layers (positive values along PC1) with shorter look-back distance (negative values along PC2)
  - Consistent with previous work suggesting that posterior temporal areas perform early-stage syntactic (and lexico-semantic) processing (Hickok & Poeppel, 2000, 2007; Flick & Pylkkänen, 2020; Murphy et al., 2022)
- IFG not strongly associated with heads specialized for particular syntactic operations despite being well-predicted by both BERT embeddings and transformations (Fig. 23B)
  - Natural language stimuli used may not contain sufficient syntactic complexity to tax IFG
  - Cortical parcellation used may yield imprecise functional localization of IFG (Fedorenko & Blank, 2020)
  - IFG may be more involved in language production than comprehension (Matchin & Hickok, 2020)

Limitations

- Pretrained BERT-base model
  - Not trained in a biologically plausible manner; allows for bi-directional information flow and has access to both past and future tokens
  - Perhaps language models with more biologically-motivated architectures and human-like objectives will provide deeper insights into human language faculty
    - Do such models exist?
- Temporal resolution of fMRI is not high enough to fully capture language processing that occurs on rapid timescales
- Current work sidesteps the acoustic and prosodic features of natural speech
  - Subjects are exposed to audio story-telling data. Cannot quantify amount of noise caused by irrelevant activity to even judge precision of current work?
  - Using movies as stimulus would have similar issue; how important is the quantification of error?

Future Work

- Training “bottlenecked” Transformer models that successively reduce the dimensionality of linguistic representations
  - Produce more hierarchical embeddings
  - Provide better structural mapping onto cortical language circuits
- May benefit from models that extract high-level contextual semantic content directly from speech signal
  - Easier said than done... Other possible methods of information isolation?